Bioinformática

da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

Apoio:

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

Elaboração de imagens

Pablo Ricardo Arantes
pablitoarantes@gmail.com

Revisão de texto

Liana Guimarães Sachett
lianasachett@gmail.com

Conteúdos

Apresentação

A ideia deste livro surgiu a partir da minha experiência pessoal com duas disciplinas em bioinformática, uma para o curso de graduação em Biomedicina e uma para o Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia, ambos na Universidade Federal do Rio Grande do Sul.

Tanto para formação em nível de graduação quanto pós-graduação, desde cedo me deparei com uma ausência quase total de materiais didáticos em português (e nacionais!), de perfil mais geral, aplicável a cursos de graduação, com poucas e importantes excessões, que devem ser mencionadas pelo seu papel pioneiro, dentre as quais destaco:

MORGON, Nelson H.; COUTINHO, K. ***Métodos de Química Teórica e Modelagem Molecular***. São Paulo: Editora Livraria da Física, 2007.

MIR, Luis ***Genômica***. São Paulo: Atheneu, 2004.

À primeira vista, química teórica e bioinformática são assuntos sem correlação. E, de fato, as pesquisas nestas áreas "puras" frequentemente apresentam pouca ou nenhuma sobreposição. De um lado, temos o estudo das propriedades estruturais e eletrônicas de moléculas e, de outro, o estudo de sequências de nucleotídeos, aminoácidos e a busca por assinalamento de funções a estas sequências. Há, assim, uma aparente separação entre, por exemplo, campos de força e árvores Bayesianas. Contudo, esta separação é apenas aparente, tendo em vista que a manifestação da função gênica passa por estruturas tridimensionais de biomoléculas. Um polimorfismo de nucleotídeo único acarreta em uma mudança na conformação e dinâmica de uma proteína, o que por sua vez pode interferir em sua função. Por outro lado, a flexibilidade de regiões de proteínas pode muitas vezes ser relacionada a eventos evolutivos, ampliando nosso entendimento do sistema em estudo e permitindo, assim, a realização de extrapolações a sistemas ortólogos ou parálogos.

Assim, **Bioinformática: da Biologia à Flexibilidade Molecular** *emprega uma definição abrangente para bioinformática, envolvendo qualquer técnica computacional aplicada ao estudo de sistemas biológicos (como o próprio nome sugere). Busca, por conseguinte, oferecer uma percepção multidisciplinar (ou talvez já estejamos beirando a transdisciplinaridade?) da área, abordando tanto aspectos relacionados a sequências de nucleotídeos e aminoácidos quanto a estrutura e dinâmica de proteínas. Adicionalmente, considerando que técnicas experimentais baseadas no uso de computadores devem, idealmente, ter seus resultados comparados a técnicas experimentais não-computacionais, este livro também inclui capítulos com algumas das técnicas experimentais mais frequentemente empregadas na validação dos números que os programas nos oferecem.*

Nesta visão, de certa forma holística, buscamos abordar não somente ácidos nucleicos e proteínas, mas carboidratos e membranas biológicas. À exceção do último, todos são agrupados como biopolímeros buscando facilitar a construção de relações entre monômeros formadores, suas conexões e as características dos polímeros resultantes. Afinal de contas, todas as células possuem membranas, e 2/3 das proteí-

nas de eucariotos são glicosiladas. Assim, busca-se oferecer ao leitor uma percepção mais próxima da importância de todas estas biomoléculas para a vida e, em muitos casos, sua participação em processos patológicos.

A linguagem escolhida para este material foi focada nas áreas biológicas e da saúde, tendo em vista que estas compreendem talvez o maior volume de problemas alvo abordados por estas técnicas. Adicionalmente, destaque foi dado na aplicação das ferramentas em detrimento do esmiuçamento de teoria, códigos, metodologias e implementações, para as quais um grande número de livros mais avançados e específicos está disponível. Em contrapartida, esta linguagem pode contribuir para que alunos de cursos de áreas não-biológicas visualizem o problema por um foco distinto, aproximando-os assim do problema alvo.

Cada capítulo foi portanto organizado com um foco principal na formação em Bioinformática para cursos de graduação. Há, contudo, diversas inserções ao longo do texto, em vermelho e fonte diferente, que buscam oferecer detalhes mais avançados, potencialmente úteis a alunos de pós-graduação. Ao final, a definição dos conceitos-chave de cada capítulo foi incluída. Tal foco na graduação nos levou a maximizar a tradução de expressões do inglês para o português, mencionando sempre a expressão inglesa original, para fins de referência. Contudo, em vários casos, a amplitude do uso de expressões originadas no inglês nos levou a mantê-las no texto, pois a tradução não teria eco nas demais fontes de leitura na área. Outra escolha envolveu a omissão de endereços na web, em decorrência de sua frequente modificação. Contudo, a partir do nome das ferramentas, não deve haver dificuldades para que os leitores identifiquem-nas pelos buscadores comuns na internet.

Embora tenhamos nos dedicado a empregar uma linguagem geral e acessível, creio que este esforço estivesse fadado a ser incompleto desde seu início em decorrência da amplitude de áreas que compõe a bioinformática. Assim, alguns capítulos serão de leitura mais fácil para alunos de cursos com maior formação em bioquímica, outros em biologia molecular, ou ainda em programação. Vejo este esforço de construção de uma linguagem comum para a área como uma obra em constante desenvolvimento e, caso o material seja de proveito para vocês, certamente nos dedicaremos a evoluí-lo em uma próxima edição.

Todo o livro foi organizado para ser aproveitado de forma digital, principalmente em tablets*. Fontes maiores foram empregadas para que a leitura fosse mais fácil e menos cansativa nestas telas. E a distribuição do material, gratuita, para um acesso o mais democrático possível entre os estudantes.*

Por fim, ao esperar que estes megabytes *de texto e fotos possam lhe ser úteis, contribuindo para sua aproximação à bioinformática, quiçá incentive-os a se aprofundarem na área, agradeço a todos os que contribuiram para a elaboração deste material. Sem eles, seu tempo, dedicação, excelência e experiência, todo este esforço não seria possível.*

Hugo Verli

Autores

Bruno César Feltes
Centro de Biotecnologia, UFRGS

Camila S. de Magalhães
Pólo de Xerém, UFRJ

Charley Christian Staats
Centro de Biotecnologia, UFRGS

Dennis Maletich Junqueira
Depto Genética, UFRGS

Diego Bonatto
Centro de Biotecnologia, UFRGS

Edwin A. Yates
Instituto de Biologia Integrativa, Universidade de Liverpool

Fabio Lima Custódio
Laboratório Nacional de Computação Científica

Fernanda Rabaioli da Silva
Centro de Biotecnologia, UFRGS

Fernando V. Maluf
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Glaucius Oliva
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Gregório K. Rocha
Laboratório Nacional de Computação Científica

Guilherme Loss de Morais
Laboratório Nacional de Computação Científica

Helena B. Nader
Departamento de Bioquímica, Unifesp

Hugo Verli
Centro de Biotecnologia, UFRGS

Isabella A. Guedes
Laboratório Nacional de Computação Científica

Ivarne L. S. Tersariol
Departamento de Bioquímica, Unifesp

João Renato C. Muniz
Grupo de Biotecnologia Molecular, IFSC - USP

Joice de Faria Poloni
Centro de Biotecnologia, UFRGS

Laurent E. Dardenne
Laboratório Nacional de Computação Científica

Luís Maurício T. R. Lima
Faculdade de Farmácia, UFRJ

Marcelo A. Lima
Departamento de Bioquímica, Unifesp

Marcius da Silva Almeida
Instituto de Bioquímica Médica, UFRJ

Priscila V. S. Z. Capriles
PPG Modelagem Computacional, UFJF

Raphael Trevizani
Laboratório Nacional de Computação Científica

Rafael V. C. Guido
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Rodrigo Ligabue Braun
Centro de Biotecnologia, UFRGS

Rogério Margis
Centro de Biotecnologia, UFRGS

Yraima Cordeiro
Faculdade de Farmácia, UFRJ

Agradecimentos

O esforço de elaboração deste livro não seria possível sem a dedicação de todos os autores. Por isso agradeço inicialmente a todos que contribuíram para este material e acreditaram na proposta de um material gratuito e digital, em sua origem. Tal esforço implicou em meses de trabalho gratuito, para o benefício dos alunos.

Agradeço especificamente ao Pablo, Rodrigo e Liana que, gastaram incontáveis horas na elaboração de figuras e revisão do texto.

Este livro é fruto da excelência acadêmica de seus autores, originada de anos dedicados à atividade científica no mais alto nível. E tal atividade só foi possível através do fomento de órgão como CNPq, CAPES, FAPERGS, FAPESP e FAPERJ aos quais, em nome de todos os autores, agradeço.

Este reconhecimento se estende às Universidades e Institutos de Pesquisa nas quais os autores estão sediados, com seus apoios físicos, logísticos, administrativos e financeiros. Nominalmente, estas instituições incluem: UFRGS, UFRJ, Universidade de Liverpool, LNCC, Unifesp, IFSC-USP e UFJF.

1. O que é Bioinformática?

"O todo sem a parte não é todo,
A parte sem o todo não é parte,
Mas se a parte o faz todo, sendo parte,
Não se diga, que é parte, sendo todo."

Gregório de Matos Guerra (1636-1696)

Hugo Verli

1.1. Introdução

Gregório de Matos, poeta brasileiro que viveu no século XVII, há quase 400 anos apresentou, na frase de epígrafe deste capítulo, seu entendimento sobre a indissociabilidade das partes para compreensão do todo. No nosso caso, o todo é a bioinformática. As partes, contudo, não são tão óbvias quanto se possa imaginar em um primeiro momento. Tampouco há consenso sobre estas. Assim, nossa discussão sobre o que é bioinformática não pretende estabelecer definições rígidas, mas guias para que o leitor entenda o quão complexa e dinâmica é esta jovem ciência.

Esta complexidade usualmente nos passa despercebida. Por exemplo, quando pensamos no impacto do projeto genoma humano, uma das principais implicações é a melhoria dos processos terapêuticos acessíveis à população. Mas a identificação de um novo gene ou mutação em um gene conhecido, por mais que seja associado a um processo patológico, está a uma grande distância de um novo fármaco. A partir da sequência, o paradigma mais moderno para desenvolvimento de novos fármacos passa pela caracterização da estrutura tridimensional da proteína codificada. Esta estrutura é então empregada para guiar o planejamento racional de novos compostos, como se um chaveiro construísse uma chave (o fármaco) a partir da fechadura. Por mais que a analogia seja simples, ainda serve como base para algumas das mais frequentes estratégias de planejamento de fármacos. E, embora a ideia de que este processo é flexível, e não rígido (mais como uma mão encaixando em uma luva, sendo a mão o fármaco e a luva o receptor) date da década de 1960, são processos tão complexos que demoramos em torno de 15 anos para lançar um novo fármaco no mercado (e este tempo não está diminuindo).

Assim, ao invés de procurar definições restritivas, este livro se propõe a empregar definições amplas, que sirvam de suporte para um entendimendo da grande gama de potencialidades e aplicações da bioinformática, buscando suportar inclusive futuras aplicações da metodologia, ainda em desenvolvimento ou por serem desenvolvidas.

Ao mesmo tempo que sequências codificantes geram seus efeitos biológicos como estruturas tridimensionais, o estudo destas pode e muito se beneficiar do estudo de sequências de proteínas relacionadas (por exemplo, alças flexíveis tendem a apresentar uma elevada variabilidade filogenética). Mesmo o estudo de sequências não codificantes pode se beneficiar do conhecimento de estruturas tridimensionais, visto que a regulação de sua expressão é realizada por fatores de transcrição proteicos. Assim, há uma retroalimentação entre as informações originadas em sequências biológicas e em suas respectivas estruturas 3D.

Em linhas gerais, este livro parte do entendimento de que a bioinformática se refere

ao emprego de ferramentas computacionais no estudo de problemas e questões biológicas, abrangendo também as aplicações relacionadas à saúde humana como o planejamento de novos fármacos.

Neste caminho, da sequência de nucleotídeos até estruturas proteicas, alcançando por fim fármacos, diversas áreas do conhecimento estão envolvidas. Biologia molecular, biologia celular, bioquímica, química, física e computação são talvez as principais grandes áreas do saber envolvidas nesse processo, cada uma contribuindo com diversas especialidades.

1.2. Origens

O que apresentaremos neste livro como bioinformática pode ser separado em duas grandes vertentes:

i) a bioinformática tradicional, ou clássica (pela primazia do nome bioinformática), que aborda principalmente problemas relacionados a sequências de nucleotídeos e aminoácidos, e

ii) a bioinformática estrutural, que aborda questões biológicas de um ponto de vista tridimensional, abrangendo a maior parte das técnicas compreendidas pela química computacional ou modelagem molecular.

Podemos traçar como momento chave para ambas as vertentes da bioinformática o início da década de 1950, quando a revista *Nature* publicou o trabalho clássico sobre a estrutura em hélice da molécula de DNA por James Watson e Francis Crick (Figura 1-1). Neste momento, as bases moleculares para o entendimento estrutural da replicação e tradução do material genético foram apresentadas, permitindo-nos entender como aquela "sequência de letras" (as bases do DNA) se organizam tridimensionalmente.

Este trabalho, contudo, deve ser visto como parte de um momento histórico, composto por diversas contribuições fundamentais para o nosso entendimento de moléculas biológicas e suas funções. Dentre estas des-

Figura 1-1: Watson e Crick em frente a um modelo da hélice de DNA. Cavendish Laboratory, Universidade de Cambridge, 1953, reproduzida sob licença.

tacam-se os trabalhos de Linus Pauling e Robert Corey, no início da década de 1950, e de Gopalasamudram N. Ramachandran, no início da década de 1960, que ofereceram as bases para a compreensão da estrutura tridimensional de proteínas.

Desde estes trabalhos até a primeira vez em que se relatou o uso de programas de computadores para visualizar estruturas tridimensionais de moléculas passaram-se mais de 10 anos quando, em 1966, Cyrus Levinthal publica na revista *Scientific American* o trabalho desenvolvido no *Massachusetts Institute of Technology* por John Ward e Robert Stotz.

Ainda nesta década se dá o primeiro esforço de sistematização do conhecimento acerca da estrutura tridimensional dos efetores da informação genética, as proteínas, em 1965, com o *Atlas of Protein Sequence and Structure*, organizado por diversos autores, dentre os quais destacaremos Margaret Dayhoff.

Este destaque se deve ao fato do papel-chave exercido pela Dra. Dayhoff na formação das raízes do que entendemos hoje por

bioinformática, tanto em sua faceta voltada para sequências quanto para estruturas. Foi uma das pioneiras no uso de computadores para o estudo de biomoléculas, incluindo tanto ácidos nucleicos quanto proteínas. Por exemplo, é ela que inicia o uso da representação de uma única letra para descrever cada aminoácido (Tabela 1-1), ao invés das usuais três letras, em uma época em que os dados eram armazenados em cartões perfurados (Figura 2-1). Desenvolveu as primeiras matrizes de substituição e fez importantes contribuições no desenvolvimento dos estudos filogenéticos. Também teve participação importante no desenvolvimento de métodos para o estudo de moléculas por cristalografia de raios-X (como veremos no capítulo 13).

Com o desenvolvimento de computadores mais poderosos e com o avanço no entendimento dos determinantes da estrutura e da dinâmica proteica, tornam-se possíveis os primeiros estudos acerca da dinâmica e do enovelamento de proteínas por simulações de dinâmica molecular por Michael Levitt e Arieh Warshel, nos anos de 1970, estudos estes agraciados com o prêmio Nobel de Química em 2013 (Figura 3-1).

A partir dos trabalhos destes e de outros pesquisadores, diversos avanços foram feitos progressivamente nos anos que se seguiram, tanto no entendimento de biomoléculas quanto no emprego de técnicas computacionais para retroalimentar este entendimento. Por exemplo, o aumento na obtenção de informações de alta qualidade sobre a estrutura 3D de biomoléculas vem servindo de suporte para o desenvolvimento de campos de força cada vez mais precisos, enquanto novas abordagens vêm possibilitando o alinhamento de sequências cada vez mais distantes evolutivamente.

Contudo talvez possamos afirmar que, a partir destas bases, os maiores impactos da área na ciência estejam se delineando neste exato período da história, em que dois importantes fatores se manifestam: o avanço (e barateamento) no poder computacional e os projetos genoma.

Computadores cada vez mais rápidos e

Tabela 1-1: Nomes dos 20 aminoácidos codificadores de proteínas junto a suas representações em 1 e 3 letras.

Aminoácido	Representação de 3 letras	Representação de 1 letra
Alanina	Ala	A
Cisteína	Cys	C
Ác. aspártico	Asp	D
Ác. glutâmico	Glu	E
Fenilalanina	Phe	F
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Lisina	Lys	K
Leucina	Leu	L
Metionina	Met	M
Asparagina	Asn	N
Prolina	Pro	P
Glutamina	Gln	Q
Arginina	Arg	R
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Triptofano	Trp	W
Tirosina	Tyr	Y

mais baratos nos permitem abordar problemas, literalmente, inimagináveis há poucos anos. Os métodos e a dimensão dos problemas abordados por um aluno de iniciação científica serão, em sua maioria, totalmente obsoletos ao final de seu doutoramento (considerado o mesmo nível de impacto dos veículos de divulgação). A cada ano que passa podemos abordar problemas mais complexos, de forma mais completa, e mais pesquisadores com menos recursos podem trabalhar nestas áreas de pesquisa, o que torna a bioinformática uma das áreas do conhecimento mais acessíveis para pesquisadores em início de carreira.

Em contrapartida, esta situação acarreta na necessidade de atualização e renovação dos procedimentos computacionais constantemente para nos mantermos competitivos na comunidade científica da área. O trabalho

Figura 2-1: IBM 7090, computador que Margaret Dayhoff utilizou no início de seus trabalhos (NASA Ames Resarch Center, 1961).

que alguém tenha publicado com simulações por dinâmica molecular (capítulo 8) alguns anos atrás, com uma simulação de, digamos, 10 ns, hoje estaria totalmente desatualizado, exigindo no mínimo uma ordem de grandeza a mais (idealmente, com replicatas e/ou condições adicionais como controle). Como consequência, as conclusões obtidas em um trabalho não necessariamente se manteriam em um novo trabalho. Similarmente, uma árvore filogenética obtida a partir de um determinado alinhamento e matriz de pontuação há 20 anos poderia ser diferente hoje, com ferramentas mais robustas de alinhamento (como será visto no capítulo 3). Esta é uma situação bastante desafiadora, assim como uma grande oportunidade, para os futuros bioinformatas.

Mas esta situação por si não é suficiente para o aumento explosivo do emprego de estratégias computacionais no estudo de sistemas biológicos, o que é principalmente devido ao projeto Genoma Humano. A partir deste, e da popularização de outros projetos genoma (capítulo 4), criou-se um gigantesco e crescente volume de sequências de genes cujas relações evolutivas e funcionais precisam ser elucidadas, como ponto de partida para novos desenvolvimentos terapêuticos. Hoje, é possível identificar um novo candidato a receptor alvo de novos fármacos a partir de organismos muito distantes evolutivamente de nós, como leveduras, bactérias ou mesmo plantas.

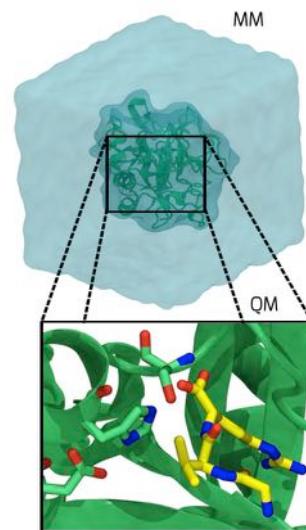



Figura 3-1: Agraciados pelo prêmio Nobel de química de 2013, os Professores Martin Karplus, Michael Levitt e Arieh Warshel.

O crescimento deste volume de informações ainda está longe de cessar. Estudos de transcriptoma, metaboloma ou glicoma ainda têm muito a agregar no nosso conhecimento do funcionamento de sistemas biológicos, potencializando tanto aplicações terapêuticas quanto biotecnológicas. Contudo, isto exigirá cada vez mais avanços da bioinformática, seja em *hardware*, *software* ou em estratégias de análise de dados e construção de modelos.

Um exemplo neste sentido envolve a gigantesca defasagem entre nossa capacidade de lidar com sequências e com estruturas 3D. Enquanto em um computador pessoal simples podemos realizar alinhamentos com algumas centenas de sequências sem maiores dificuldades, localmente ou na *web*, dependendo do método, e recebendo a resposta quase que imediatamente, para realizar uma simulação por dinâmica molecular de uma única proteína precisaríamos, neste mesmo computador, de alguns meses.

Um último aspecto importante nesta contextualização inicial da bioinformática, dentro da proposta apresentada por este livro, diz respeito à importância relativa das diferentes biomoléculas na manifestação da informação genética, mantendo a homeostasia e servindo como alvo de modulação far-

macológica ou emprego biotecnológico. Tradicionalmente, os ácidos nucleicos e as proteínas receberam a maior atenção enquanto alvos da bioinformática, os primeiros como repositórios da informação biológica e as últimas como efetores desta informação. Esta percepção, contudo, vem sendo progressivamente relativizada. Membranas e carboidratos, a despeito de não estarem codificados diretamente no genoma (não há um códon para um fosfolipídeo ou para um monossacarídeo), são fundamentais à homeostasia da grande maioria dos organismos em todos os domínios da vida. E entender estes papéis vem se tornando um importante alvo da bioinformática.

1.3. Problemas alvo

Considerando o tipo de informação manipulada, os problemas e questões abordados pela bioinformática podem ser agrupados entre aqueles relacionados a sequências de biomoléculas e aqueles relacionados à estrutura de biomoléculas (Figura 4-1). À primeira vista, considerando que de forma geral estruturas de proteínas são determinadas por seus genes, poderíamos imaginar que lidar com estruturas 3D seria redundante a manipular sequências, conjuntos de informações 1D. Esta percepção é limitada e não se configura como verdade para diversas questões. Na verdade, existem aspectos únicos em cada conjunto de informação, não diretamente transferíveis para o outro.

Inicialmente, como veremos adiante (item 1.4 e capítulo 2), o enovelamento de proteínas é um fenômeno extremamente complexo e ainda não totalmente compreendido, de forma que não somos capazes de transformar uma sequência linear de aminoácidos (codificada por seu gene) em uma estrutura 3D (salvo para algumas situações específicas, que serão vistas ao longo do livro).

Outro aspecto importante é que o enovelamento de proteínas, em muitas situações, depende de mais do que sua sequência de aminoácidos, envolvendo aspectos como o ambiente e o local onde a proteína estará na célula ou organismo, a ocorrência de modificação co- ou pós-traducionais e a sua interação com chaperonas. Para ilustrar o quanto este fenômeno é complexo, embora diversas sequências com identidade mínima possam ter estruturas 3D extremamente parecidas, em alguns casos a troca de um ou poucos resíduos de aminoácidos pode modificar totalmente a função, chegando até a interferir na forma tridimensional que uma proteína adota.

Em contrapartida, algumas informações presentes em sequências gênicas ou mesmo peptídicas não são necessariamente observáveis em estruturas tridimensionais. Por exemplo, regiões promotoras ou reguladoras da expressão gênica são facilmente descritas como informações 1D, e peptídeos sinal ou íntrons estão normalmente ausentes nas formas nativas de proteínas, sendo mais facilmente observáveis por sequências das biomoléculas em questão.

Adicionalmente, estruturas 3D de moléculas são formas muito mais complexas de serem manipuladas que sequências 1D, o que agrega uma série de dificuldades nos estudos de bioinformática. Assim, diversas tarefas tendem a ser muito simplificadas (ou mesmo de outra forma não seriam possíveis atualmente) quando trabalhamos com sequências em vez de estruturas. Por exemplo, a identificação de uma assinatura para modificação pós-traducional é muito mais ágil em uma sequência do que em um conjunto de milhares de átomos distribuídos em um espaço tridimensional.

Por fim, talvez o motivo mais prático para separarmos as duas abordagens se refere à facilidade de obtenção das informações. Os métodos experimentais para sequenciamento de ácidos nucleicos estão muito mais avançados do que os métodos para determinação da estrutura 3D de biomoléculas. A diferença de capacidade de determinação dos dois conjuntos de dados é de ordens de grandeza.

Questões relacionadas a sequências

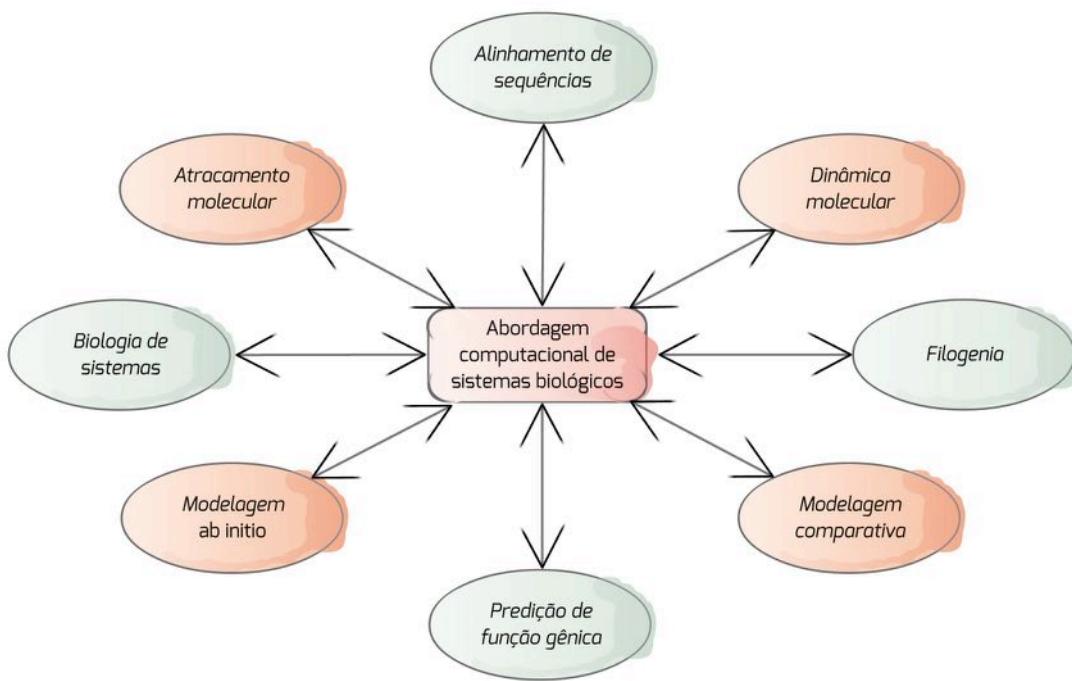



Figura 4-1: Representação de algumas das principais áreas da bioinformática. As metodologias que lidam majoritariamente com estruturas 3D estão representadas em laranja, enquanto as metodologias envolvidas principalmente com sequências estão representadas em verde. Devemos lembrar, contudo, que esta separação é imperfeita. Por exemplo, a modelagem comparativa parte de sequências, a função de um gene pode ser determinada pela estrutura da proteína associada.

A manipulação de sequências é menos custosa computacionalmente, nos possibilitando lidar com genomas inteiros. Isto permite realizar análises em indivíduos ou mesmo populações de indivíduos, nos aproximando do entendimendo dos organismos em sua complexidade biológica. Podemos traçar a história evolutiva de um conjunto de organismos ou construir redes de interação entre centenas ou milhares de moléculas de um determinado organismo, tecido ou tipo celular. Em linhas gerais, os objetos de estudo relacionados a sequências de biomoléculas incluem:

i) comparações entre sequências (alinhamento);
ii) identificação de padrões em sequências (assinaturas);
iii) caracterização de relações evolutivas (filogenia);
iv) construção e anotação de genomas;
v) construção de redes (biologia de sistemas).

Vale destacar que estas análises podem receber a contribuição de estudos envolvendo a estrutura das biomoléculas de interesse ou mesmo ser validadas por estas. Por exemplo, resíduos conservados evolutivamente possuem grande chance de possuírem papel funcional (como atuando na catálise) ou estrutural (estabilizando a estutura proteica). Assim, comparar um alinhamento à estrutura 3D pode tanto explicar quanto oferecer novas abordagens e considerações ao significado de conservações de resíduos maiores ou menores em conjuntos de sequências.

Questões relacionadas a estruturas

Ao contrário da manipulação de sequências, estruturas exigem um maior poder de processamento para serem manipuladas. Na prática, podemos manipular uma ou um pequeno punhado de estruturas simultaneamente (embora este número venha crescendo progressivamente). Neste caso, o foco costuma ser o entendimento de moléculas e dos eventos mediados por estas, individualmente, incluindo:

i) obtenção de modelos 3D para proteínas e outras biomoléculas (por exemplo, modelagem comparativa);
ii) identificação do modo de interação de moléculas (atracamento);
iii) seleção de compostos com maior potencial de inibição (atracamento);
iv) caracterização da flexibilidade molecular (dinâmica molecular);
v) avaliação do efeito de mudanças na estrutura e ambiente molecular na dinâmica e função de biomoléculas (dinâmica molecular).

O uso de sequências para alimentar estudos estruturais é mais comum na construção de modelos tridimensionais de proteínas a partir de suas sequências codificadoras, no método denominado modelagem comparativa (capítulo 7). Contudo, outras relações extremamente úteis podem ser estabelecidas. Por exemplo, por serem estruturas usualmente flexíveis, alças tendem a possuir uma maior capacidade de acomodar mutações ao longo da evolução. Isto permite uma comparação entre resultados de alinhamentos e, por exemplo, perfis de flexibilidade observáveis através de simulações por dinâmica molecular.

1.4. Tendências e desfios

Como uma área em rápido desenvolvimento, a bioinformática exige de seu praticante uma constante atenção a novas abordagens, métodos, requerimentos e tendências. Programas podem se tornar rapidamente ineficientes comparados a novas ferramentas ou mesmo obsoletos. Avanços de *hardware* podem (e na verdade vem fazendo isso) catapultar o nível de exigência metodológica pelas revistas de ponta. E há algumas áreas em específico nas quais a comunidade científica vem concentrando esforços. São por conseguinte áreas de grande impacto potencial e grande competição na literatura científica, dentre as quais destacaremos algumas abaixo.

Processamento em CPU e GPU

CPUs (*Central Processing Unit*s ou unidades de processamento central) ou simplesmente processadores (ou ainda microprocessadores) são partes dos computadores responsáveis pela execução das instruções estabelecidas pelos programas. Desde seu surgimento em torno da metade do século XX, as CPUs tornaram-se progressivamente mais complexas, confiáveis, rápidas e baratas. Esse processo foi previsto pioneiramente por Gordon E. Moore, no que ficou sendo conhecido desde então como a lei de Moore. Segundo esta lei, o número de transistores em um processador (na verdade em qualquer circuito integrado) dobra aproximadamente a cada 2 anos (Figura 5-1). O impacto do fenômeno descrito nesta observação na vida moderna é enorme, envolvendo desde nossos computadores, celulares e câmeras digitais até a precisão de estudos climáticos (com impacto na prevenção de catástrofes e na agricultura), medicina, engenharia, indústria bélica e aeroespacial. Com o aumento da velocidade e barateamento das CPUs, podemos a cada ano construir modelos mais precisos de fenômenos biológicos progressivamente mais complexos. Na prática, o avanço da bioinformática está ligado intrinsecamente à lei de Moore.

Em uma CPU podemos encontrar não somente um microprocessador, mas mais de um, o que é chamado multi-processamento e estas CPUs de processadores de múltiplos núcleos (*multi-core processing*). Hoje, a grande maioria dos processadores empregados em computadores, *notebooks* e celulares já possui múltiplos núcleos. Se o programa que estamos utilizando for adaptado para este tipo de processamento, o cálculo poderá ser distribuído pelos núcleos de processamento, tornando o cálculo significativamente mais rápido. A grande maioria dos aplicativos em bioinformática já possui versões compatíveis com processamento em múltiplos núcleos, e devemos estar atentos à escolha destas versões e à instalação de forma que essa característica esteja funcional, sob pena de subutilização da CPU.

Já GPUs (*Graphical Processing Unit*s ou unidades de processamento gráfico) são microprocessadores desenvolvidos inicialmente

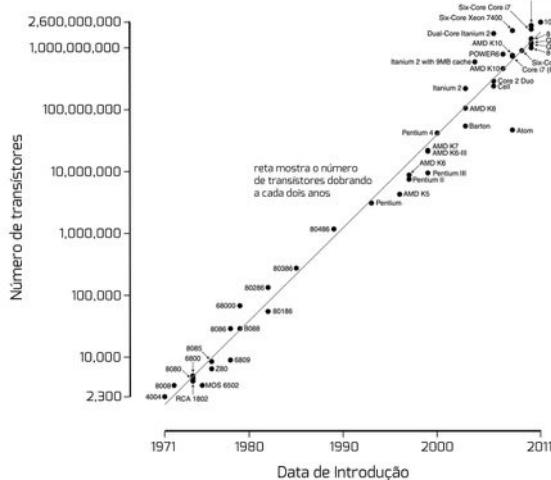



Figura 5-1: Representação da lei de Moore, indicando o aumento no número de transistores em microprocessadores no período de 1971 a 2011. Adaptada de William Wegman, 2011 (*Creative Commons*).

como unidades especializadas na manipulação de representações gráficas em computadores. Estão, assim, normalmente localizadas nas placas de vídeo de nossos computadores. O termo GPU foi popularizado a partir de 1999 com o lançamento da placa de vídeo GeForce256, comercializada pela Nvidia.

O desenvolvimento das GPUs remonta ao início dos anos de 1990, com o aumento do emprego de gráficos em 3D nos computadores e videogames. De fato, alguns dos primeiros exemplos de *hardware* dedicado ao processamento em 3D estão associados a consoles como PlayStation e Nintendo 64. Atualmente, enquanto CPUs possuem até em torno de uma dezena de núcleos de processamento, GPUs podem facilmente alcançar centenas ou mesmo milhares de núcleos de processamento, permitindo uma grande aceleração na manipulação de polígonos e formas geométricas, encontradas em aplicações 3D (como os jogos) e sua renderização (Figura 6-1). Tal aumento de performance ao dividir a carga de trabalho em um grande número de núcleos de processamento abriu um grande horizonte de possibilidades em computação científica, implicando em grande aumento na velocidade de manipulação de dados.

Diversos aplicativos em bioinformática vêm sendo portados para trabalhar com GPUs. Desde o alinhamento de sequências à filogenia, do atracamento molecular à dinâmica molecular, múltiplos pacotes estão disponíveis, tanto pagos quanto gratuitos, capazes de explorar a computação em GPU, e este número vem crescendo a cada ano, apontando para uma nova tendência na área. O usuário deve, contudo, observar seu problema alvo, pois a aceleração fornecida pela GPU dependerá das características do problema em questão e da eficiência e portabilidade do código empregado.

A combinação de CPUs e GPUs com múltiplos núcleos fez com que a capacidade de processamento de alguns supercomputadores de há alguns anos já esteja disponível para computadores pessoais, nos chamados supercomputadores pessoais.

Predições a partir de sequências

Quando estudamos uma sequência de nucleotídeos de DNA desconhecida é importante determinar seu papel funcional, por exemplo, se codificante de proteínas ou não. E, sendo codificante, qual proteína é produzida ao final da tradução e qual sua função. Tais predições são realizadas a partir de algoritmos construídos a partir de bancos de dados

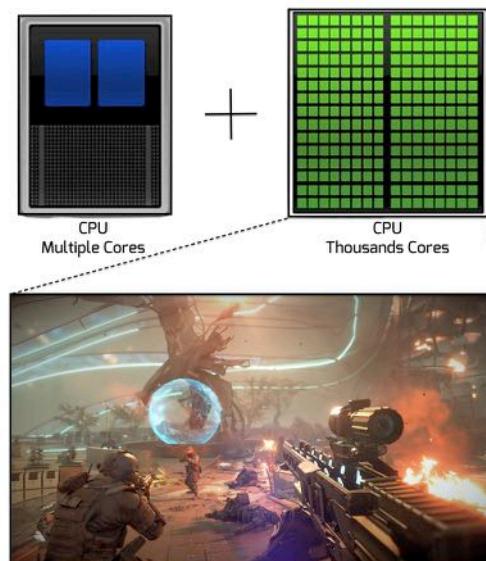



Figura 6-1: Representação dos núcleos de processamento em CPUs e GPUs. O grande número de núcleos em GPUs permite a realização de cálculos complexos rapidamente.

existentes, relacionando determinada sequência a características e propriedades específicas. Contudo, somente uma pequena quantidade de organismos teve seu genoma sequenciado até o momento e, destes, somente uma pequena parte de genes teve sua função determinada experimentalmente. Devemos, portanto, lembrar que as predições destes modelos estão relacionadas a quão completos foram os bancos de dados que os basearam. E que estes estão em contínuo avanço (ou seja, uma predição feita há 5 anos não necessariamente será igual a uma predição hoje que, por sua vez, pode ser diferente de uma predição de função gênica daqui a 5 anos - discutiremos no capítulo 3 alguns indicadores da qualidade dessas associações).

Predição de energia livre

Os fenômenos moleculares são regidos pela termodinâmica, tanto para reações químicas na síntese de um novo fármaco quanto à ação da DNA polimerase ou ao enovelamento de proteínas. Entender termos como entropia, entalpia e energia livre torna-se, assim, fundamental na adequada descrição destes fenômenos e, a partir desta, sua previsão computacional. Quando a medida destas variáveis se tornar precisa o bastante, poderemos esperar a substituição de diversos experimentos em bancada por cálculos em computadores mas, infelizmente, ainda não chegamos neste momento.

Predições de energia livre tem impacto direto na identificação da estrutura 2ária de moléculas de RNA, na localização de regiões do DNA para ligação de reguladores da transcrição, para a especificidade de enzimas por substratos e receptores por ligantes ou moduladores (fisiológicos ou terapêuticos, isto é, fármacos). Assim, diversos métodos foram desenvolvidos para a obtenção destas medidas, tais como a perturbação da energia livre, a integração termodinâmica, a energia de interação linear, a metadinâmica e diversas estratégias empíricas voltadas ao pareamento de nucleotídeos ou atracamento molecular.

A despeito desta diversidade de estratégias, a predição da energia livre em processos moleculares continua sendo um grande desafio. Em decorrência do elevado custo computacional associado a estes cálculos, diferentes tipos de simplificações e generalizações precisam ser realizadas, comprometendo nossa capacidade de empregá-los de forma ampla e fidedigna.

Enovelamento de proteínas

Como veremos adiante no livro, o enovelamento de proteínas é um dos processos mais complexos conhecidos pelo ser humano. O número de estados conformacionais possíveis para uma proteína pequena é gigantesco, dos quais um ou alguns poucos serão observáveis em solução em condições nativas. Os métodos experimentais usualmente empregados para tal, a cristalografia de raios-X e a ressonância magnética nuclear, são métodos caros e ainda possuem algumas limitações importantes em determinadas situações, apontando para a Bioinformática um potencial e importante papel na determinação da estrutura de biomoléculas.

Mas para que precisamos saber como é a estrutura tridimensional de uma determinada biomolécula? Esta pergunta possui muitas respostas, incluindo a compreensão de como a natureza evoluiu, como os organismos funcionam, como os processos patológicos se desenvolvem (e podem ser tratados) e como as enzimas exercem suas funções catalíticas. Tomemos este último caso como exemplo.

Com o entendimento de como proteínas se enovelam, será possível construir novas proteínas, capazes de adotar formas que a natureza não previu até o momento, enzimas aptas a catalizar reações de importância econômica, com menor toxicidade, o que terá por si impacto ambiental. Ainda, abre-se a possibilidade de planejamento racional de enzimas e proteínas envolvidas na detoxificação de áreas. Esta linha de pesquisa está em seu início, e o número de grupos de pesquisa dedicados ao redor do mundo para trabalhar na

engenharia de proteínas vem aumentando gradativamente. Mas, infelizmente, ainda não possuimos uma base teórica que nos permita entender e prever, com precisão e de forma ampla, a estrutura 3D de proteínas.

Contudo, esta problemática vem sendo abordada a cada ano com maior sucesso. Para proteínas com no mínimo em torno de 30% de identidade com outras proteínas de estrutura 3D já determinada, podem ser obtidos modelos de qualidade próxima àquela de métodos experimentais. Em outros casos, estruturas cristalográficas podem ser refinadas por métodos computacionais, agregando explicitamente informações ausentes nos experimentos (como a flexibilidade molecular). Outro exemplo é a construção de alças flexíveis, de difícil observação experimental mas que podem ser abordadas por diferentes métodos computacionais.

Para ácidos nucleicos, a construção computacional de estruturas 3D de moléculas de DNA é tarefa relativamente simples, que usualmente não requer os custos associados a experimentos de cristalografia e ressonância magnética. Para moléculas de RNA, contudo, a elevada flexibilidade traz consigo desafios adicionais. Mesmo assim, em diversos casos as estratégias computacionais possuem vantagens em lidar com moléculas muito flexíveis. Talvez o caso mais emblemático neste sentido sejam as membranas biológicas. Estas macromoléculas biológicas não são observáveis nos experimentos usuais capazes de determinar estruturas com resolução atômica, embora através de simulações por dinâmica molecular tenham suas estruturas descritas com elevada fidelidade.

Outro caso em que os métodos computacionais parecem possuir vantagens em relação aos experimentais envolve os carboidratos. Embora sejam moléculas em vários aspectos mais complexos que proteínas, carboidratos biológicos não parecem sofrer enovelamento nem adotar tipos de estrutura $2^{ária}$ em solução (embora o façam em ambiente cristalino), o que os torna na prática um problema estrutural mais simples que proteínas. De fato, vem sendo possível prever a estrutura de glicanas com graus variados de complexidade com grande precisão, um campo no qual os métodos experimentais possuem grandes dificuldades em abordar.

Validação experimental

Em linhas gerais, métodos computacionais devem ser comparados a dados experimentais para validação. Esta afirmação, embora tomada geralmente como um axioma, é bastante simplista, e não expressa claramente a complexidade e desafio nesta tarefa. Alguns pontos específicos incluem:

i) nem sempre há dados experimentais disponíveis para validar os cálculos e simulações realizados. Por exemplo, este é o caso com frequência para alinhamentos de sequências, para relações filogenéticas, para predições *ab initio* da estrutura de proteínas e para a descrição da flexibilidade de biomoléculas obtidas por dinâmica molecular. Nem sempre há fósseis ou outras evidências arqueológicas para validar antepassados evidenciados por estudos filogenéticos. Por outro lado, não há métodos experimentais com resolução atômica e temporal, de forma que a validação de simulações por dinâmica molecular é em grande medida indireta (uma estrutura obtida por cristalografia é única, sem variação temporal, enquanto os modelos oriundos de ressonância magnética nuclear correspondem a médias durante o período de coleta do dado);

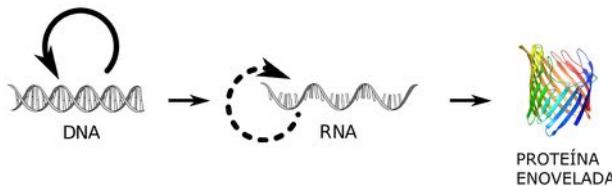
ii) os dados experimentais devem ser adequados ao estudo computacional empregado. Assim, se estamos estudando a formação de um complexo fármaco-receptor, resultados *in vivo* devem ser evitados, enquanto os experimentos *in vitro* preferidos. Se administramos um determinado fármaco por via oral a um camundongo, este fármaco passará por diversos processos farmacocinéticos (absorção, distribuição, metabolização e excreção) que muito provavelmente irão interferir na ação

frente ao receptor alvo. Portanto, para estudos de atracamento, dados *in vivo* devem ser evitados;

iii) a margem de erro do dado experimental deve ser considerada quando comparada aos dados computacionais. Frequentemente a margem de erro para experimentos na bancada é maior que para aqueles realizados em computadores, limitando a extensão da validação. Usando novamente o exemplo de estudos de atracamento, se a afinidade experimental de um fármaco por seu receptor é de 0,11 ± 0,04 μM, valores teóricos de 97 nM a 105 nM estarão corretos. Por outro lado, frequentemente os resultados experimentais são expressos como a menor dose testada, por exemplo, > 5 μM. Assim, qualquer valor maior que 5 μM será validado pelo dado experimental, o que cria uma grande dificuldade de validação (como comparar 5 a, digamos, 1.000?);

iv) as condições nas quais os experimentos foram realizadas devem ser observadas com estrito cuidado. Temperatura, contaminantes, sais e concentrações diferentes daquelas no ambiente nativo são frequentemente requeridas por alguns métodos experimentais, e podem interferir nos resultados. Por exemplo, a melitina (principal componente do veneno da abelha *Apis mellifera*) aparece como uma hélice em estudos cristalográficos mas é desenovelada no plasma humano, como pode ser confirmado por experimentos de dicroismo circular com força iônica compatível com o plasma.

Assim, a despeito do axioma da exigência de validação experimental para estudos computacionais, não é infrequente que um dado computacional apresente maior precisão que um dado obtido na bancada. Na realidade, um modelo computacional, frequentemente chamado de teórico em oposição aos métodos ditos experimentais, não é nada além de um experimento computacional que, infelizmente, nem sempre tem contraparte em experimentos de "bancada". E esses adjetivos não carregam consigo qualificações quanto à confiabilidade dos resultados gerados.

1.5. Leitura recomendada

KHATRI, Purvesh; DRAGHICI, Sorin. Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. ***Bioinformatics***, 21, 3587-3593, 2005.

MORGON, Nelson H.; COUTINHO, K. ***Métodos de Química Teórica e Modelagem Molecular***. São Paulo: Editora Livraria da Física, 2007.

MIR, Luis. ***Genômica***. São Paulo: Atheneu, 2004.

2. Níveis de Informação Biológica

Representação do fluxo de informação em sistemas biológicos.

Hugo Verli

2.1. Introdução

Por mais que possam apresentar enormes diferenças em suas características os seres vivos, desde bactérias a mamíferos, passando por plantas e fungos, são compostos aproximadamente pelos mesmos tipos de moléculas. Estes compostos incluem proteínas, ácidos nucleicos, lipídeos e carboidratos, moléculas nas quais a vida como conhecemos é baseada.

Cada uma destas classes de biomoléculas apresenta, contudo, enormes variações de forma, estrutura e função na natureza, o que possibilita a gigantesca variedade e complexidade de manifestações da vida em nosso planeta. Mesmo em estruturas que não são normalmente consideradas vivas, como é o caso dos vírus, estas biomoléculas são também encontradas e se mostram essenciais à execução de suas funções, sejam estas patológicas ou não.

Independentemente da forma pela qual a vida se manifesta, a informação que a rege está armazenada nas moléculas de DNA. Contudo, tais dados não são usados diretamente, mas através de uma molécula intermediária, o RNA (mais precisamente o RNAm), sintetizado por um processo denominado transcrição (uma molécula de ácido nucleico é transcrita em outra molécula de ácido nucleico). Esta molécula de RNAm irá servir como molde para a síntese de proteínas, em um processo chamado de tradução (uma molécula de ácido nucleico é traduzida em uma molécula de proteína). As proteínas, assim expressas, irão reger a maioria dos fenômenos relacionados à função dos organismos e à perpetuação da vida (embora diversos outros processos sejam modulados por outras biomoléculas). Esta informação segue um sentido tão conservado na natureza que foi convencionado denominá-lo como dogma central da biologia molecular (Figura 1-2).

A importância do dogma central no entendimento da informação e função biológicas pode ser exemplificada no fato de que ele aborda os três tipos mais comuns de moléculas estudadas por técnicas de bioinformática, o DNA, o RNA e as proteínas, estabelecendo um fluxo de informação universal à vida como conhecemos. Adicionalmente, a efetivação da informação genética, através das proteínas, acarreta na construção e manutenção de outras biomoléculas, igualmente essenciais ao desenvolvimento da vida, como carboidratos e lipídeos. Em decorrência de sua elevada massa molecular, proteínas, ácidos nucleicos, lipídeos agregados em membranas e carboidratos complexos são chamados de macromoléculas.

Embora carboidratos e lipídeos não estejam explici-

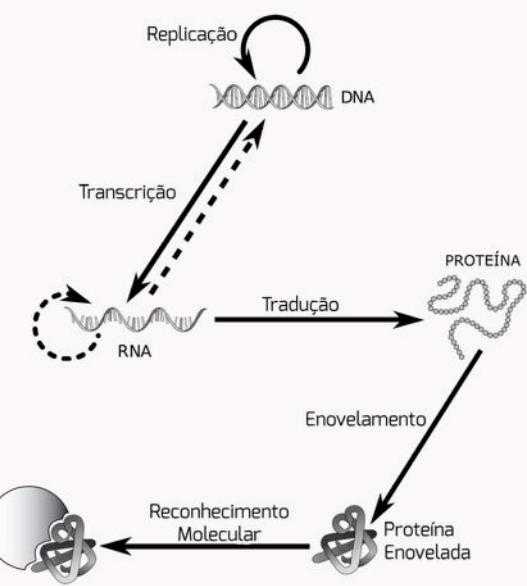



Figura 1-2: Representação do dogma central da biologia molecular, no qual o fluxo de informação em sistemas biológicos é descrito, desde seu armazenamento no DNA até a manifestação da função biológica. O esquema tradicional sofreu a adição do processo de enovelamento de de reconhecimento molecular devido ao seu caráter fundamental para a manifestação da função gênica. Adaptado de Hupé, 2012.

tamente inseridos no dogma central, não devemos minimizar sua importância. Apesar de por muito tempo estes compostos terem sido reconhecidos simplesmente por papéis energéticos e estruturais, ambos vêm sendo demonstrados como envolvidos em inúmeros fenômenos biológicos, como na glicosilação de proteínas e na formação de jangadas lipídicas. Estes, por sua vez, podem interferir diretamente na execução da função de proteínas e na homeostasia dos organismos.

Não somente macromoléculas são importantes biologicamente. Proteínas sintetizam uma infinidade de compostos de baixa massa molecular, ou micromoléculas, que atuam como neurotransmissores, sinalizadores e moduladores dos mais variados tipos representando, portanto, diferentes tipos de informação em sistemas biológicos. Por exemplo, a infecção do nosso organismo por bactérias desencadeia um processo inflamatório mediado por derivados lipídicos denominados prostaglandinas. Para combater micro-organismos competidores, fungos e bactérias produzem pequenos compostos com atividade antibiótica, muitos destes usados até hoje como fármacos. Desta forma, se a bioinformática se dedica ao estudo, por ferramentas computacionais, dos fenômenos relacionados à vida, o estudo de micromoléculas também torna-se foco da bioinformática ao abordar compostos relacionados à manutenção fisiológica ou terapêutica (neste caso, no planejamento de novos candidatos a agentes terapêuticos).

As técnicas modernas de bioinformática são capazes de lidar com todas estas biomoléculas que, contudo, possuem particularidades derivadas de suas diferenças químicas. Tais aspectos devem ser conhecidos de forma a permitir a construção de modelos computacionais mais precisos e adequados ao estudo dos mais diversos aspectos relacionados à vida.

Não há uma forma única de representar as diferentes moléculas biológicas. Cada estratégia de representação possui suas vantagens e desvantagens, que devem ser avaliadas de acordo com o estudo em andamento. Estratégias com menor volume de informação associado possuem menor custo computacional e, portanto, nos permitem avaliar rapidamente grandes quantidades de dados, por exemplo, genomas inteiros de diferentes organismos, cada um contendo dezenas de milhares de proteínas. Por outro lado, estratégias com maior volume de informação associado acarretam em custo computacional gigantesco nos limitando a, por exemplo, um punhado de proteínas, de dois ou três organismos. O trânsito por tal disparidade é um dos grandes desafios atuais para o profissional que trabalha com bioinformática.

2.2. Macromoléculas biológicas

As biomoléculas descritas no dogma central da biologia molecular, proteínas, DNA e RNA, são o que chamamos de biopolímeros, isto é, polímeros produzidos pelos seres vivos. Somam-se a este grupo de moléculas os carboidratos, que também podem ser encontrados como polímeros em meio biológico.

As propriedades de um polímero tornam-se consequência das propriedades de suas unidades monoméricas constituintes. No

caso dos biopolímeros, os monômeros podem ser aminoácidos, nucleotídeos e monossacarídeos. Assim, o conhecimento destas unidades básicas irá auxiliar diretamente no estudo de suas formas poliméricas e, por conseguinte, das funções biológicas destes polímeros sintetizados na natureza.

Ácidos nucleicos

Os compostos denominados ácidos nucleicos são polímeros sintetizados a partir de unidades denominadas nucleotídeos. Os nucleotídeos são formados por três partes constituintes: uma base nitrogenada, um carboidrato e um grupo fosfato. A base nitrogenada pode ser adenina (A), guanina (G), citosina (C), uracila (U) ou timina (T), enquanto a parte sacarídica poderá ser β-D-ribose (frequentemente abreviada simplesmente como ribose, para o RNA) ou a 2-desoxi-β-D-ribose (usualmente abreviada como desoxirribose, para o DNA) (Figura 2-2). Nas moléculas de ácidos nucleicos, os nucleotídeos são ligados através da denominada ligação fosfodiéster (ver adiante).

Quando a base nitrogenada está ligada ao carboidrato, na ausência do grupo fosfato, os compostos gerados são denominados nucleosídeos. Formados por ligação de diferentes nucleotídeos à β-D-ribose temos a

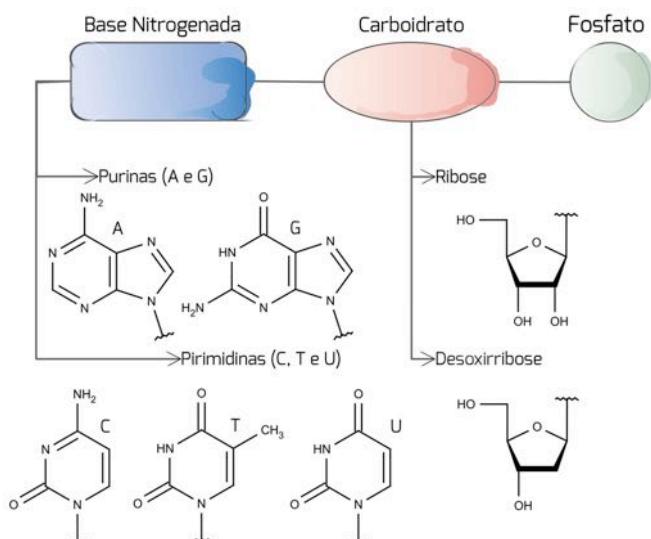



Figura 2-2: Representação esquemática de um nucleotídeo e suas variações na base nitrogenada e no carboidrato.

adenosina, a guanosina, a citidina, a uridina e a timidina. A estes compostos podem ainda se ligar diferentes números de grupos fosfato. Assim, a adenosina pode se apresentar monofosfatada (AMP, do inglês *adenosine monophosphate*), difosfatada (ADP, do inglês *adenosine diphosphate*) ou ainda trifosfatada (ATP, do inglês *adenosine triphosphate*).

Conforme veremos adiante, carboidratos apresentam características conformacionais específicas, como sua capacidade de deformar seu anel em diferentes estados conformacionais. Esta característica se soma à grande flexibilidade da ligação fostodiéster na criação de um esqueleto bastante flexível para ácidos nucleicos. Em contrapartida a esta flexibilidade da parte sacarídica dos nucleotídeos, cada base nitrogenada é essencialmente planar, uma vez que constituem-se de anéis aromáticos, e portanto apresentam flexibilidade bastante reduzida.

Proteínas

As proteínas são polímeros sintetizados pelas células a partir de aminoácidos. São talvez as biomoléculas mais versáteis na natureza, sendo capazes de adotar uma gigantesca possibilidade de arranjos tridimensionais, não encontrada nos demais biopolímeros. Não por acaso, constituem-se no principal produto direto da informação genética, a partir da tradução do RNAm.

O genoma codifica diretamente 20 aminoácidos (22 contando selenocisteína e pirrolisina, que são codificadas por codons de parada) para composição de proteínas (Figura 3-2), embora outros resíduos de aminoácidos, não codificados no genoma (Figura 4-2), possam ser sintetizados a partir destes e exercer funções bastante específicas, como o ácido γ-amino butírico (GABA), um neurotransmissor inibitório no sistema nervoso central, ou como o resíduo ácido γ-carbóxi glutâmico (GLA), constituinte de diversas proteínas plasmáticas e fundamental na hemostasia.

Os aminoácidos codificados no genoma apresentam algumas características bem definidas e compartilhadas entre si. Todos os resíduos apresentam uma região comum, independente do resíduo. Esta região é denomi-

Representação 2D de um resíduo de aminoácido dentro de uma sequência polipeptídica

Representação 3D de um resíduo de aminoácido dentro de uma sequência polipeptídica

Resíduos polares

Resíduos apolares

não-carregados

alifáticos

aromáticos

carregados

negativos

positivos

Aspartato, Asp ou D

Glutamato, Glu ou E

Arginina, Arg ou R

Lisina, Lys ou K

Serina, Ser ou S

Treonina, Thr ou T

Cisteína, Cys ou C

Metionina, Met ou M

Asparagina, Asn ou N

Glutamina, Gln ou Q

Glicina, Gly ou G

Alanina, Ala ou A

Valina, Val ou V

Leucina, Leu ou L

Isoleucina, Ile ou I

Prolina, Pro ou

Fenilalanina, Phe ou F

Tirosina, Tyr ou Y

Histidina, His ou H

Triptofano, Trp ou W

Figura 3-2: Estrutura dos aminoácidos codificados no genoma, organizados segundo as propriedades de suas cadeias laterais. No topo o esqueleto peptídico é representado como encontrado dentro de uma proteína, tanto em sua forma 2D quanto 3D. Nesta última, o grupo R (cadeia lateral) está apresentado como uma esfera amarela, enquanto a continuação da cadeia polipeptídica como esferas verde-escuras. As cadeias laterais estão apresentadas em sua ionização mais comum, plasmática.

nada esqueleto peptídico, e é composta pelo grupo amino, pelo grupo ácido carboxílico e pelo átomo de carbono que liga estes dois grupos, denominado carbono α (Cα). A diferença entre estes resíduos está no grupamento ligado ao Cα, chamado cadeia lateral (Figura 3-2).

Enantiômeros são compostos que, diferindo somente no arranjo de seus átomos no espaço (como no caso de L-Ser e D-Ser), correspondem um à imagem especular do outro (isto é, uma é o reflexo em um espelho da outra).

À exceção da glicina, todos os aminoácidos são quirais, em decorrência da presença de quatro substiuintes diferentes ligados ao Cα. Salvo casos específicos, todos os aminoácidos quirais são encontrados em somente uma forma enantiomérica, L. Como consequência, todas as proteínas são quirais, e isto tem implicações importantes em fenômenos bioquímicos e na prática terapêutica.

Dois enantiômeros interagem de forma idêntica com compostos que não sejam quirais. Por exemplo, a

5-Hidroxilisina
6-*N*-metil lisina
Selenocisteína
γ-Carbóxiglutamato
4-Hidroxiprolina

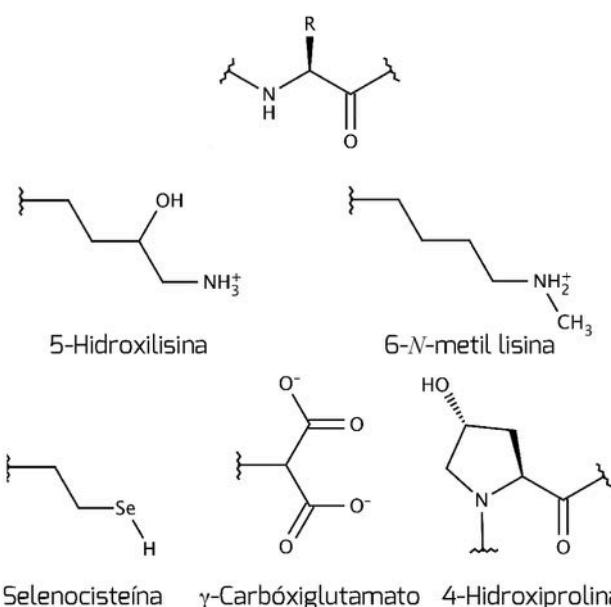



Figura 4-2: Exemplos de aminoácidos encontrados em nosso organismo mas não codificados no genoma humano.

interação de L-Ser e D-Ser com a água é idêntica. Em contrapartida, compostos quirais interagem diferentemente com cada enantiômero. Assim, a interação de L-Ser e D-Ser com uma dada proteína seria diferente. Assim, se tivermos um fármaco quiral, uma de suas formas enantioméricas será ativa e a outra provavelmente inativa, menos ativa ou mesmo tóxica.

O esqueleto peptídico de aminoácidos apresenta um grupo do tipo ácido carboxílico somente em aminoácidos livres, monoméricos, ou na posição terminal da proteína, denominada região C-terminal (o final da sequência polipeptídica). Da mesma forma, só encontramos o grupo amino na região demominada N-terminal (o início da sequência polipeptídica). À exceção destas extremidades, os grupos amino e carboxílico reagem, dando origem a um grupo amida. Assim, dentro de uma proteína, cada aminoácido contribui com um um átomo de nitrogênio e com uma carbonila para a formação de uma amida contida no esqueleto peptídico.

Os aminoácidos frequentemente são agrupados de acordo com as propriedades de suas cadeias laterais (Figura 3-2). Inicialmente, podem ser separados em resíduos polares e apolares. Os resíduos polares incluem aminoácidos não-carregados e carregados (com carga positiva ou negativa), enquanto os resíduos apolares incluem aminoácidos aromáticos e alifáticos (não aromáticos).

As propriedades dos aminoácidos são altamente influenciadas pelo pH do meio circundante. De acordo com sua acidez ou basicidade, a carga dos resíduos pode ser modificada e, por conseguinte, algumas propriedades da proteína. Assim, dependendo do compartimento celular, uma mesma proteína pode apresentar ionização distinta de seus resíduos de aminoácidos e, por conseguinte, propriedades eletrostáticas diferentes. Tais características destacam a importância de uma avaliação adequada do estado de ionização dos resíduos de aminoácidos das proteínas em estudo, principalmente o resíduo de histidina.

Durante a síntese proteica, os aminoácidos são conectados através da denominada ligação peptídica (ver adiante). Neste processo, o grupo carboxilato de um resíduo e o o grupo amino de outro resíduo de aminoácido reagem, dando origem a um grupo amida que compõe a ligação peptídica.

Carboidratos

Carboidratos compõem um terceiro grupo de biomoléculas. São compostos que, ao contrário das proteínas, não estão codificados diretamente no genoma. Enquanto a síntese de proteínas é guiada por um molde (a molécula de RNAm), a síntese de carboidratos não segue uma referência direta, mas um processo complexo e menos específico.

Embora o genoma não codifique a sequência oligossacarídica, ele determina a expressão de diversas enzimas que sintetizam carboidratos, ligam-os a outras estruturas polissacarídicas ou ainda modificam os resíduos monossacarídicos, adicionando ou removendo grupamentos substituintes nos anéis furanosídicos ou piranosídicos (Figura 5-2). Todo este processo é bastante específico, envolvendo tipos de monossacarídeos ou ainda posições específicas dentro destas moléculas. Uma das principais famílias de enzimas envolvidas neste processo são as denominadas glicosil transferases.

Esta família de biomoléculas apresenta uma grande variedade de formas (e, por conseguinte, funções), desde suas formas monoméricas até grandes polímeros com centenas de unidades monossacarídicas. São encontrados ligados a proteínas, formando as chamadas glicoproteínas; sulfatados, dando origem aos glicosaminoglicanos; ligados a lipídeos em membranas celulares (os glicolipí-

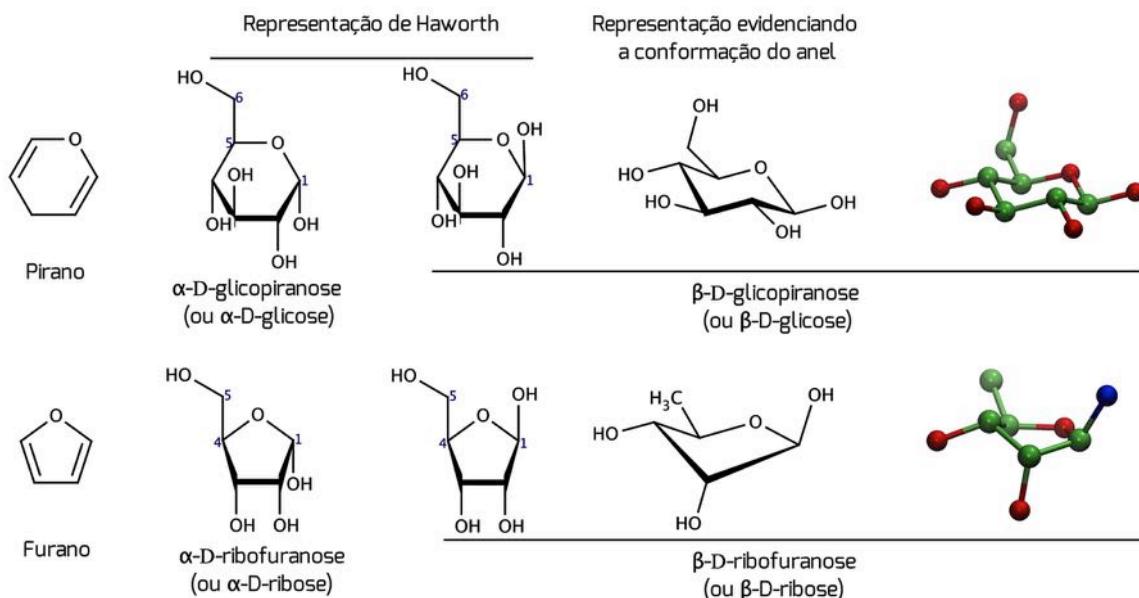



Figura 5-2: Os dois principais grupos de carboidratos envolvem monossacarídeos compostos por anéis de 5 (furanoses) e 6 membros (piranoses). São apresentados 3 tipos de visualização para estas moléculas, duas 2D e uma 3D.

deos) e como exopolissacarídeos da parede celular de fungos, dentro outros.

A forma majoritária de monossacarídeos biológicos em solução é um ciclo, mais comumente composto por 5 ou 6 átomos. Os carboidratos com anéis de 5 membros são denominados furanoses (como a ribose e a desoxirribose), por semelhança ao composto furano, enquanto os carboidratos com anéis de 6 membros são denominados piranoses (como a glicose, a manose e a galactose), pela sua similaridade com o composto pirano (Figura 5-2).

Estes anéis apresentam características conformacionais importantes. No caso das furanoses, podem ser as formas em envelope e torcida. No caso das piranoses, podem ser as formas em cadeira e bote torcido (Figura 6-2). Cada uma destas formas pode apresentar ainda variações, específicas para cada carboidrato em solução. Esta transição entre diversos estados conformacionais de monossacarídeos é denominada de equilíbrio pseudo-rotacional.

Os carboidratos possuem algumas diferenças importantes em relação aos aminoácidos. São, em geral, compostos mais polares, o que indica que irão interagir fortemente com a água. Outra diferença importante se refere à sua diversidade. Em comparação aos 20 aminoácidos codificados no genoma, mais de 100 possíveis unidades monossacarídicas já foram observadas como presentes em biomoléculas (Figura 7-2).

Em analogia à ligação peptídica, carboidratos são ligados entre si (ou a outras moléculas) através da denominada ligação glicosídica. Contudo, aminoácidos possuem somente um grupo amino e um grupo ácido carboxílico em seu esqueleto peptídico, de forma que somente um tipo de ligação peptídica é possível entre dois resíduos (o mesmo se dá com nucleotídeos). Como a ligação glicosídica entre dois monossacarídeos é formada pela reação entre dois grupos hidroximetileno (CHOH), e cada monossacarídeo possui vários destes grupos, múltiplas ligações entre dois monossacarídeos consecutivos tornam-se possíveis. Cria-se, assim, um complexo espectro de possíveis ligações entre os mesmos dois monossacarídeos.

O átomo de carbono na posição 1 (C1) de um monossacarídeo apresenta propriedades específicas, sen-

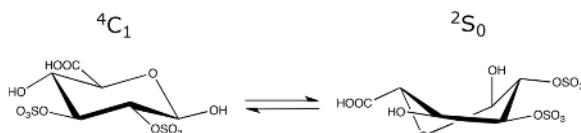



Figura 6-2: Equilíbrio conformacional entre a forma de cadeira e bote torcido para o resíduo de ácido idurônico, componente da heparina.

α-D-glicose α-D-manose α-D-fucose

α-D-*N*-acetilgalactosamina α-L-iduronato ácido α-L-neuramínico

α-D-bacilosamina 6-*C*-sulfo-α-D-fucose 2NAc-2,4-dideoxi-hex-5-ulose

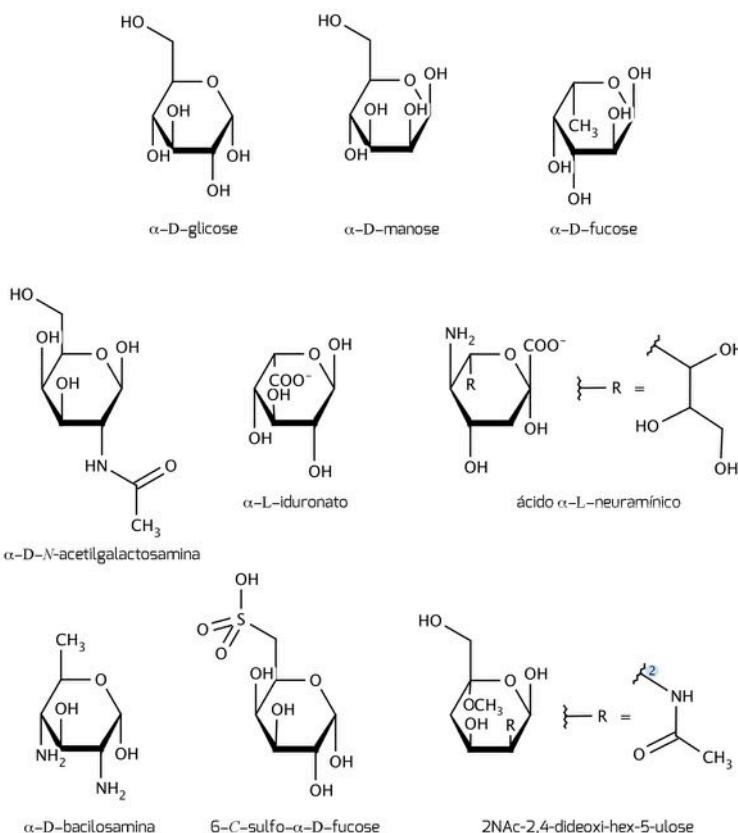



Figura 7-2: Exemplo da complexidade de possíveis monossacarídeos encontrados na natureza.

do denominado carbono anomérico. Para um mesmo monossacarídeo, o carbono anomérico pode ser encontrado em duas possíveis configurações, α e β (Figura 5-2). Assim, uma ligação glicosídica entre o carbono anomérico (C1) de uma manose e o átomo C3 de outra manose poderia ocorrer de duas formas, α-Man-(1→3)-Man ou β-Man-(1→3)-Man. No caso de glicoproteínas, contudo, a forma α é aquela usualmente encontrada para o resíduo de manose (para outros resíduos, a forma anomérica preferencial pode ser diferente).

Tomando como exemplo o tetrassacarídeo α-Man-(1→2)-α-Man-(1→2)-α-Man-(1→3)-Man, comumente encontrado em glicoproteínas do tipo oligomanose, o primeiro resíduo de manose (denominada extremidade não-redutora) possui seu carbono anomérico ocupado na ligação glicosídica, tendo sua configuração (neste exemplo α) fixa. Em contrapartida, o quarto resíduo de manose possui seu carbono anomérico livre. Esta porção é denominada redutora, e tem a configuração do carbono anomérico variável, isto é, pode estar tanto na forma α quanto β.

Membranas

Diferentemente dos ácidos nucleicos, proteínas e carboidratos, membranas não se constituem em polímeros biológicos, mas em agregados moleculares de lipídeos anfipáticos organizando uma bicamada (Figura 8-2). Apresentam papel fundamental à vida, compartimentalizando a célula, definindo seus limites, propriedades e organizando estruturas celulares.

É importante ter em mente que membranas são muito mais do que simples "paredes" delimitadoras da célula. Os componentes de membranas são variados, incluídos diferentes tipos de lipídeos, proteínas e carboidratos. A presença e localização destes componentes pode ser modulada de forma dinâmica em função de necessidades da célula, tecido ou organismo, sinalizando e modulando cadeias de eventos e definindo regiões da célula com propriedades específicas (a chamada polaridade celular).

Moléculas anfipáticas apresentam como característica a presença simultânea de uma região polar, também chamada de cabeça polar (hidrofílica ou lipofóbica) e de uma região apolar, também chamada de cauda hidrofóbica (hidrofóbica ou lipofílica). Assim, membranas celulares possuem superfícies polares e

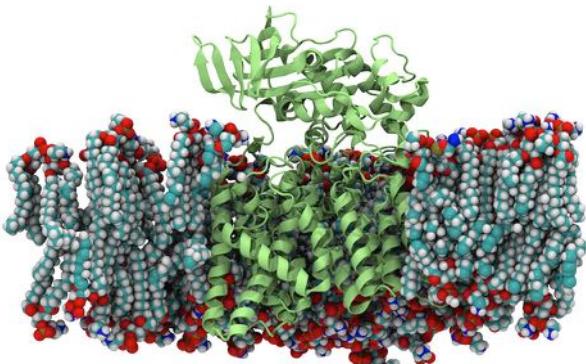



Figura 8-2: Representação de uma membrana POPE (palmitoil oleil fosfatidil etanolamina) contendo a enzima PglB (oligossacaril transferase) de *Campylobacter lari*. Os átomos de oxigênio estão representados em vermelho, os átomos de carbono em verde, os átomos de hidrogênio em branco e nitrogênios em azul. A enzima está representada como cartoon verde.

interiores apolares. As características destas duas regiões, contudo, podem variar bastante em função da composição dos lipídeos, interferindo na carga, espessura e fluidez da membrana (e, por conseguinte, na sua capacidade de modular fenômenos biológicos).

"Micromoléculas" biológicas

Quando pensamos nos efetores da informação genética é natural que a primeira família de biomoléculas que venha a nossa mente seja a das proteínas, codificadas diretamente no genoma. Contudo, como vimos anteriormente, outros tipos de biomoléculas são fundamentais ao funcionamento dos organismos, mesmo que estas não estejam codificadas diretamente no DNA.

Da mesma forma como não há um conjunto de bases nitrogenadas que codifique monossacarídeos ou lipídeos, diversos compostos de baixa massa molecular (por isso muitas vezes chamados de micromoléculas, em oposição às macromoléculas, compostos de elevada massa molecular) não possuem codificação direta no genoma, mas são produzidos a partir de enzimas que, estas sim, têm suas sequências de aminoácidos definidas pela molécula de DNA. Neurotransmissores, hormônios, metabólitos primários e secundários em plantas e uma infinidade de compostos, em decorrência de sua importância biológica (e terapêutica), são potenciais alvos de estudos computacionais. Contudo, justamente em decorrência de sua grande variedade química, torna-se difícil estabelecer padrões ou referências estruturais, como é o caso das biomacromoléculas vistas anteriormente. Frequentemente, esta característica cria uma série de dificuldades e desafios no emprego de ferramentas computacionais no estudo de micromoléculas. Dentre estas dificuldades destaca-se a necessidade de desenvolvimento de parâmetros específicos para cada molécula (como veremos no capítulo 8).

2.3. Níveis de organização

A classificação da estrutura de biomacromoléculas envolve, didaticamente, quatro diferentes níveis de complexidade. Esta separação facilita o nosso entendimento do como e do porquê macromoléculas adotarem determinadas formas em meio biológico e, a partir destas, desempenharem funções específicas. Adicionalmente, cada nível traz volume e tipos de informação diferentes, exigindo poder computacional e abordagens distintas, como veremos adiante.

Em princípio, estes níveis apresentam um componente hierárquico, ou seja, a informação de um nível é importante ou necessária para o nível de complexidade seguinte. Contudo, outros fatores podem participar neste processo.

Por exemplo, no caso das proteínas, embora normalmente consideremos que a informação contida na estrutura 1ária (isto é, a sua sequência de aminoácidos) seja determinante para a sua estrutura 2ária, ela não é o único determinante. Concessões podem ser realizadas para permitir uma estrutra 3ária ou mesmo 4ária mais estável.

Assim, uma determinada região em hélice pode ser parcialmente desestruturada para facilitar a formação de um determinado domínio (ver adiante). Este tipo de consideração é importante na validação de modelos teóricos para a estrutura de proteínas, como veremos no capítulo 7.

Adicionalmente, fatores externos à própria sequência proteica podem interferir nestes níveis de organização. Um dos fatores mais comuns é a glicosilação de proteínas, que frequentemente estabiliza partes da mesma e, assim como as chaperonas, pode interferir na forma proteica tridimensional existente em meio biológico.

*Estrutura 1ª^ria^*

O nível inicial de complexidade, a estrutura 1ária, consiste num padrão de letras (ou pequenos conjuntos de letras) que representa a composição do biopolímero. Esta sequência de letras representa uma informação de natureza unidimensional (1D), em que a única dimensão descrita é a ordem de aparecimento dos monômeros.

Para ácidos nucleicos, a estrutura 1ária consiste numa sequência de nucleotídeos, enquanto para proteínas em uma sequência de aminoácidos e, para carboidratos, em uma sequência de monossacarídeos (Figura 9-2). Este último caso é o único para o qual não há uma descrição de uma única letra para cada monômero, principalmente em face do elevado número de possíveis monômeros encontrados na natureza, maior que o número de letras no alfabeto.

Embora de menor complexidade, a estrutura 1ária nos oferece um grande volume de informações sobre a forma nativa da biomolécula e, por conseguinte, sobre suas funções. Tais informações advém principalmente da comparação de sequências de biomoléculas (aminoácidos ou nucleotídeos) em busca de padrões específicos associados a determinadas características ou funções. Uma vez identificados, esses padrões ou assinaturas podem ser usados na busca das mesmas características em outras proteínas, desconhecidas. Estas comparações ainda nos permitem estudar a evolução destas biomoléculas e de seus organismos, contribuindo no entendimento de como a vida se desenvolveu e atingiu o seu estágio atual de complexidade (ver capítulo 5).

DNA:
GGTATAGGCGCTGTTCTTAAGGTGCTAACAACGGGGT
TACCCGCGTTGATCTCGTGGATAAAACGCAAACGCCA
ACAG

RNA:
GGUAUAGGCGCUGUUCUUAAGGUGCUAACAACGGG
GUUACCCGCGUUGAUCUCGUGGAUAAAACGCAAAC
GCCAACAG

Aminoácidos:
GIGAVLKVLTTGLPALISWIKRKRQQ

Sequência sacarídica:
α-D-GlcNAc,6S-(1→3)-β-D-GlcA-(1→4)-α-D-GlcNS,3S,6S-(1→4)-α-L-IdoA,2S-(1→4)-α-D-GlcNS,6S

Figura 9-2: Representação da estrutura 1ária de diferentes biomacromoléculas: DNA, RNA, proteína (estas três representando o peptídeo melitina, componente do veneno da abelha *Apis mellifera*) e carboidratos (representando uma sequência repetitiva de heparina). A letra S na sequência oligossacarídica indica sulfatação.

Estrutura 2ária

A partir da sequência de monômeros descritos, em uma determinada ordem específica, na estutura 1ária surgem interações entre monômeros vizinhos e com as moléculas de solvente circundantes. Por exemplo, enquanto dois nucleotídeos vizinhos tendem a "empilhar" os anéis das bases, uma cadeia lateral de um aminoácido polar vai se expor à água, maximizando interações por ligação de hidrogênio com este solvente. De forma semelhante, uma cadeia apolar irá se expor aos lipídeos em uma membrana, maximizando interações hidrofóbicas com este outro solvente.

Estas interações entre monômeros acabam por dar origem a padrões repetitivos de organização espacial, denominados de estrutura 2ária (Figura 10-2). Estes padrões ou elementos aparecem em número relativa-

mente pequeno de tipos, de forma que a estrutura tridimensional de biomoléculas pode ser descrita como uma combinação de conjuntos destes elementos.

Diferentes composições de estrutura 1ária podem gerar um mesmo tipo de estrutura 2ária. Não por acaso, as propriedades destas estruturas 2árias, mesmo que formadas por sequências diferentes, apresentam semelhanças. Por exemplo, uma alça em proteínas é frequentemente uma estrutura 2ária bastante flexível, enquanto folhas e hélices tendem a ser mais rígidas.

As estuturas 2árias mais frequentemente lembradas são aquelas relacionadas a proteínas. Incluem três grupos de elementos principais: as alças, as hélices e as folhas β.

As alças ou voltas são elementos envolvidos na conexão entre hélices e folhas. Tendem a ser, portanto, estruturas flexíveis para acomodar as mais variadas orientações que estas hélices e fitas podem adotar entre si. Embora alças pequenas possam ser bastante rígidas, suas flexibilidades tendem a aumentar conforme o tamanho da alça aumenta (Tabela 1-2). Justamente em função desta elevada flexibilidade, alças são mais susceptíveis evolutivamente a sofrerem mutações (salvo se estiverem sob alguma pressão evolutiva, determinada por alguma função específica). Em outras palavras, a troca de um resíduo por outro de propriedades distintas pode ser mais facilmente acomodada nesta estrutra flexível do que nos outros tipos de estrutura 2ária, mais rígidos.

Enquanto hélices e folhas apresentam periodicidade ao longo de suas estruturas (semelhança nos pares de ângulos φ e ψ a cada aminoácido, ver adiante), alças se distinguem por não apresentarem periodicidade. Ainda, embora alças sejam frequentemente consideradas como elementos sem estrutura definida (as chamadas *random coils*), ou mesmo com estrutura aleatória, isto não é sempre verdade. Alças podem adotar formas mais definidas, dependendo de seu tamanho e composição.

De forma semelhante, é equivocado subestimar a importância das alças, considerando somente seu papel como elemento de conexão. Alças apresentam diversos impactos funcionais importantes em proteínas.

Tabela 1-2: Tipos de alças mais comuns encontrados em proteínas.

Tipo	Tamanho (nº de resíduos)
voltas γ	3
voltas β	4
voltas α	5
voltas π	6
alças Ω	6-16[a]
alças ζ	6-16[a]

[a] A despeito de tamanhos semelhantes, as formas destas alças se aproximam das letras que as denominam. Na volta Ω os resíduos das extremidades da alça estão próximos, e na volta ζ observa-se uma distorção na geometria.

Por exemplo, sua flexibilidade permite que atuem como tampas ou abas, cobrindo sítios ativos e regulando o acesso de moduladores ou substratos. De forma ainda mais direta, alças são frequentemente os elementos de estrutura 2ária mais expostos ao solvente. Assim, muitas vezes envolvem-se em contatos proteína-proteína (ou com outras biomoléculas), os quais podem ser determinantes para a função proteica. Assim, embora mais susceptíveis evolutivamente a mutações, não são incomuns alças com resíduos conservados, fundamentais para suas respectivas funções biológicas.

A hélice α e as folhas β foram inicialmente descritos por Linus Pauling e Robert B. Corey em 1951, embora as primeiras propostas para as estruturas em folhas datem de décadas mais cedo, em 1933, por Astbury e Bell. As folhas β são formadas por sequências de aminoácidos (cada sequência é denominada de fita) quase completamente extendidas. Estas fitas, quase lineares, interagem lado a lado ao longo de seus eixos longitudinais, através de uma série de ligações de hidrogênio entre o grupamento N-H de uma fita e o grupamento C=O da fita vizinha (Figura 10-2). Para que esta organização seja possível, os átomos de Cα adotam orientação intercalada, acima e abaixo do plano da folha. Esta organização se assemelha a uma série de dobraduras em uma folha de papel, de forma que este tipo de estrutura 2ária é tam-

bém denominado de folhas β pregueadas (Figura 10-2).

A forma pregueada de folhas β também é acompanhada pelas cadeias laterais dos resíduos de aminoácidos, ora acima do plano da folha, ora abaixo. Contudo, resíduos em fitas vizinhas orientam suas cadeias laterais para o mesmo lado, frequentemente de forma justaposta (Figura 10-2). Isto permite, por exemplo, que uma face da folha seja hidrofóbica e a outra hidrofílica.

A organização das fitas em folhas pode seguir duas orientações possíveis: *i*) a porção N-terminal de uma fita interagindo com a porção N-terminal da fita vizinha (e, consequentemente, o C-terminal interagindo com o C-terminal), ou *ii*) a porção N-terminal de uma fita interagindo com a porção C-terminal da fita vizinha. Estas duas possibilidades de interações de fitas dão origem a dois tipos de folhas β: as paralelas e as antiparalelas.

As folhas β paralelas e antiparalelas diferem em outras características. Esta organização diferenciada das fitas acarreta, por exemplo, em um padrão distinto de ligações de hidrogênio. Enquanto nas folhas antiparalelas as ligações de hidrogênio formam um ângulo de 90º com as fitas, nas folhas paralelas estes ângulos se tornam maiores (e as interações mais fracas) (Figura 10-2).

As folhas β podem ser encontradas em formas puras, paralelas ou antiparalelas, ou mistas, em que folhas paralelas pareiam com folhas antiparalelas. Contudo, folhas β paralelas tendem a ser menos estáveis conformacionalmente que folhas β antiparalelas. Esta diferença pode ser bastante significativa, suficiente para acarretar na desnaturação de proteínas por seus inibidores, como foi proposto na ação de serpinas sob suas proteses alvo.

O trabalho pioneiro de Pauling e Corey no início dos anos 50 do século XX identificou não somente as folhas, mas também hélices em sequências polipeptídicas. A formação da hélice, de forma similar às folhas, também envolve a realização de ligações de hidrogênio entre grupos N-H e C=O vizinhos no espaço (mas não na sequência) (Figura 10-2). Contudo, enquanto nas folhas β estas interações se dão com resíduos em fitas vizinhas, nas hélices estas interações acontecem com resíduos mais próximos na sequência, entre as voltas da hélice.

Diversos tipos de hélices podem ser encontrados em proteínas (Tabela 2-2). A hélice mais comum, denominada de hélice α, apresenta 3,6 resíduos de aminoácidos por volta da hélice, e cada aminoácido (n) realiza ligação de hidrogênio com o quarto resíduo seguinte ($n + 4$), que perfaz (aproximadamente) uma volta completa da hélice. Outro tipo de hélice comum em alguns tipos de proteína é a hélice de poli-prolina II encontrada, por exemplo, em proteínas de parede celular de plantas e no colágeno. Neste tipo de hélice, contudo, como o átomo de nitrogênio da prolina está ligado a três átomos de carbono, não há formação de ligação de hidrogênio durante a organização da hélice.

Existem, ainda, outros tipos de hélice, menos comuns, como a hélice π e a hélice 3_{10} (Tabela 2-2). Quanto à nomenclatura, a hélice 3_{10} foge ao padrão de uso de letras gregas das hélices α e π. O número 3 representa o número de resíduos por volta da hélice, enquanto o número 10 reflete o número de átomos entre duas ligações de hidrogênio vizinhas dentro da hélice. Assim, segundo esta nomenclatura, a hélice α seria chamada de $3{,}6_{13}$ e a hélice π de $4{,}4_{16}$. Tais nomenclaturas, contudo, não são normalmente empregadas.

Não são só as proteínas que que apresentam estruturas $2^{árias}$. Ácidos nucleicos e carboidratos também podem apresentar padrões repetitivos de organização espacial, definidos pela sequência de monômeros que os constituem.

A molécula de DNA pode adotar três tipos de estrutura $2^{ária}$, denominados A, B e Z (Figura 11-2), embora a forma B seja a estrutura mais comum e a partir dela sejam definidas as fendas maior e menor do DNA (Tabela 3-2). A transição entre estas formas é determinada pela hidratação, tipos de cátions e da própria sequência de nucleotídeos. Contudo, a dificuldade em mimetizar as interações biológicas, envolvidas no DNA e em complexos DNA-proteínas, durante a determinação de estruturas 3D dificulta associações mais claras de cada tipo de estrutura $2^{ária}$ a fenômenos específicos *in vivo*.

Diferentes tipos de estrutura $2^{ária}$ acarretam em diferentes propriedades estruturais

A) hélice α B) hélice 3_{10} C) hélice π

D) folha β paralela E) folha β antiparalela

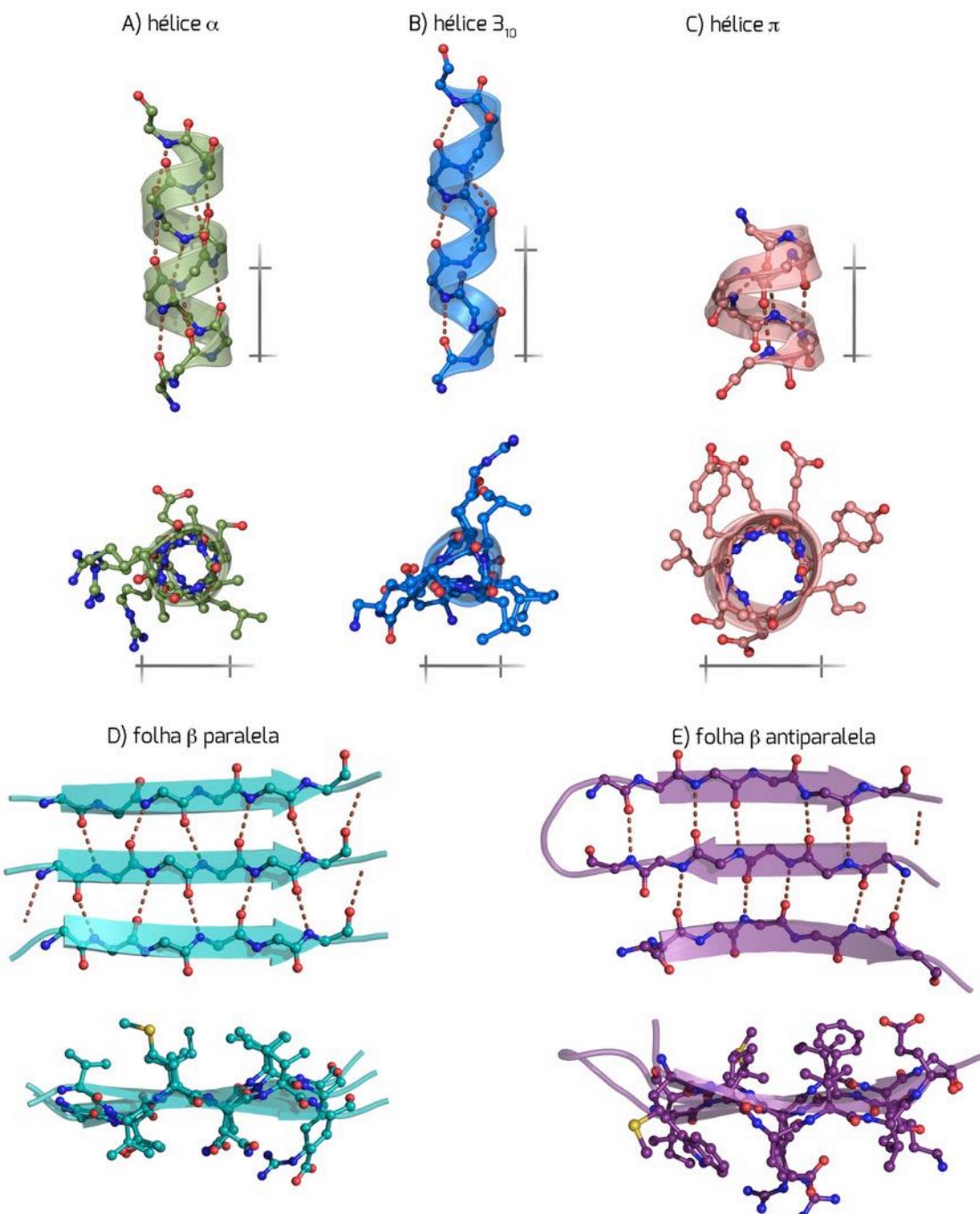



Figura 10-2: Representação dos tipos mais comuns de estrutura 2ária encontrados em proteínas. Em verde estão as hélices α (A), em azul as hélices 3_{10} (B), em salmão as hélices π (C), em ciano as folhas β paralelas (D) e roxo as antiparalelas (E). As ligações de hidrogênio entre átomos do esqueleto peptídico estão apresentadas como linhas tracejadas em marrom. As estruturas são partes que compõe as proteínas descritas pelos códigos PDB 1BD8, 1ABB, 2QD1, 1EE6 e 1PC0, e para cada uma duas diferentes orientações são apresentadas. Note que as cadeias laterais apontam para fora do eixo das hélices e, para as folhas, para cima e para baixo do plano definido pelas fitas.

na molécula de DNA, como na largura e profundidade das fendas maior e menor e na disposição e orientação dos grupos fosfato, propriedades estas que, por sua vez, estão diretamente relacionadas à especificidade da interação do DNA com proteínas e fármacos.

A forma B do DNA pode assumir dois sub-estados, denominados BI e BII, definidos por diferenças em tor-

Tabela 2-2: Tipos de hélices encontrados em proteínas.

Tipo de hélice	Resíduos / volta	Ligação de hidrogênio	Elevação / resíduo (Å)	Elevação / volta (Å)	Direção mais comum
hélice α	3,6	$n + 4$	1,5	5,4	direita
hélice 3_{10}	3	$n + 3$	2,0	6,0	direita
hélice π	4,4	$n + 5$	1,2	5,3	direita
poli-Pro I	3,3	-	1,7	5,6	direita
poli-Pro II	3	-	3,1	9,3	esquerda

ções na parte sacarídica e no grupo fosfato (ver adiante). Essa região, formada por carboidrato e fosfato, é também denominada de esqueleto do DNA, em analogia ao esqueleto peptídico. A lógica é a mesma: o esqueleto é composto pela região comum a todos os monômeros formadores do biopolímero. Adicionalmente, outras formas de DNA já foram identificadas (alguns autores afirmam inclusive que poucas letras do alfabeto sobram para nomear novas formas de DNA que por ventura venham a ser identificadas), embora muitas ainda não tenham papel biológico claro.

A maioria dos genomas eucarióticos está sujeita a um fenômeno de metilação do DNA, que consiste na adição de um grupo metila no átomo de carbono na posição 5 dos resíduos de citosina. Como uma modificação estrutural epigenética envolvida na regulação do potencial regulatório e transcricional do DNA, deve-se estar atento à necessidade de incluir tal modificação na descrição deste ácido nucleico.

Não somente o DNA, mas também o RNA possui estrutura $2^{\text{ária}}$. Contudo, ao contrário do DNA, que é uma molécula contendo duas fítas de ácidos nucleicos, na maioria das situações o RNA é uma molécula composta por uma única fita. Assim, enquanto no DNA os pareamentos entre bases que dão origem à estrutura $2^{\text{ária}}$ surgem da interação de moléculas (fitas) diferentes e complementares, no RNA a estutura $2^{\text{ária}}$ surge de interações na própria fita, que dobra-se sobre si mesma.

As estruturas $2^{\text{árias}}$ de RNA incluem regiões de bases pareadas, alças de grampos, alças internas, bojos (do inglês *bulge*) e junções. Quando o RNA se dobra sobre si, ele forma pareamentos entre bases complementares de forma análoga àquelas vistas no DNA. Quando uma das fitas no RNA pareado apresenta bases que não possuem uma contrapartida para formar um par A-U ou C-G, forma-se uma protuberância ou bojo.

Estes bojos, isto é, bases não pareadas em uma dupla-fita, também podem ser encontradas em folhas β. Neste caso, resíduos de aminoácidos de uma fita deixam de interagir com a fita vizinha, dando origem a este outro tipo de estrutura $2^{\text{ária}}$ de proteínas.

As alças de grampos em moléculas de RNA são análogas às voltas observadas em proteínas, conectando duas fitas β por um pequeno segmento de poucos resíduos. No RNA, quando a fita dobra-se sobre si mesma, deixa alguns resíduos (no mínimo 4) projetados para fora, formando uma alça. Neste tipo de estrutura $2^{\text{ária}}$, a alça está vizinha a somente uma região de pareamento de bases, enquanto que há duas regiões, a cada lado do bojo, de bases pareadas.

As alças internas podem ser entendidas como uma dupla fita de DNA em que, no seu meio, as bases não são complementares e, por isso, não pareiam. Assim, ambas as fitas apresentam bases que não estão pareadas, o que a diferencia do bojo. Por fim, as junções conectam 3 ou mais regiões de bases pareadas.

O terceiro tipo de biopolímero constituinte de biomacromoléculas, os carboidratos podem, similarmente a proteínas e ácidos nucleicos, adotar padrões repetitivos de organização de suas unidades formadoras, monossacarídeos, isto é, em elementos de estrutura $2^{\text{ária}}$.

Polissacarídeos lineares desenvolvem estruturas de hélices, similarmente à proteínas e ácidos nucleicos. No caso destas moléculas, contudo, a variabilidade de organizações possíveis é muito maior, de for-

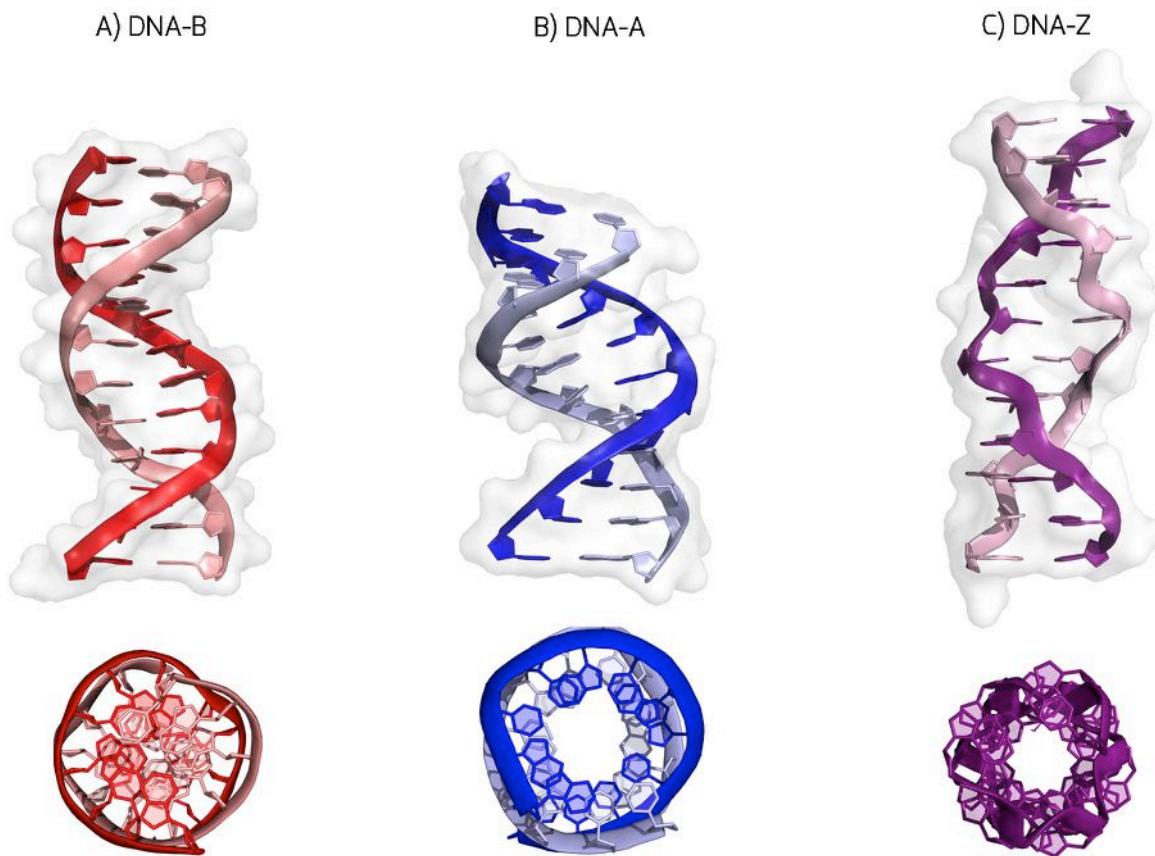



Figura 11-2: Representação dos tipos mais comuns de estrutura 2ária encontrados no DNA, ilustradas para sequências de 12 nucleotídeos. Em vermelho estão as hélices B (A), em azul as hélices A (B) e em magenta as hélices Z (C). As estruturas pelos códigos PDB 3BSE, 3V9D e 279D. Para cada uma duas diferentes orientações são apresentadas, e o esqueleto das moléculas de DNA está representado como fitas.

ma que não há definição específica para um ou alguns tipos de hélices, como vimos anteriormente. Ao invés disto, cada tipo de polissacarídeo apresentará um número de resíduos por volta, elevação por resíduo e elevação por volta, assim como seu sentido para a direita ou para a esquerda (vide tabela 2-3).

Estas características, contudo, são normalmente determinadas experimentalmente através de difração de raios-X, na qual a amostra está na fase cristalina. Esta é uma condição adequada à descrição, por exemplo, da quitina, polissacarídeo encontrado na natureza em condições semelhantes. Contudo, quando estes polissacarídeos são transpostos para soluções biológicas, estas moléculas adotam uma elevada flexibilidade e, por conseguinte, grande variação conformacional. Não raramente, perdemos a capacidade de identificar formas repetitivas, e a denominação de alças desordenadas pode também ser aplicada a polissacarídeos.

Adicionalmente, carboidratos não se apresentam somente como polissacarídeos lineares, mas como oligo- ou polissacarídeos ramificados. Esta ramificação agrega um grau adicional de complexidade na descrição da forma destes compostos. Mesmo assim, ainda é possível descrever a forma destes compostos, caso a caso, como veremos adiante.

Estrutura 3ária

A importância do conhecimento da estrutura 2ária de biomoléculas reside, principalmente, no fato de que estes elementos se organizam no espaço tridimensional, dando

Tabela 2-3: Tipos de hélices encontrados em ácidos nucleicos.

Tipo de hélice	pb / volta	Elevação / pb (Å)	Elevação / volta (Å)	Fenda maior (Å)		Fenda menor (Å)		Direção
				Largura	Profundidade	Largura	Profundidade	
DNA A	11	2,9	32	2,7	13,5	11,0	2,8	direita
DNA B	10	3,4	34	11,7	8,5	5,7	7,5	direita
DNA Z	12	3,8	45	-	convexa	4	9	esquerda

origem ao que chamamos de estrutura 3ª. Em outras palavras, a estrutura 3ª de uma dada biomolécula corresponde à montagem dos seus elementos de estrutura 2ª. Por outro lado, é a estrutura 3ª (ou a 4ª, que veremos a seguir) que irá exercer a função biológica da molécula em questão.

Os diversos elementos de estrutura 2ª de uma dada molécula se organizam em uma estrutura 3ª através de um fenômeno denominado enovelamento (também chamado em português de dobramento, do termo em inglês *folding*). Neste processo, uma combinação de forças converge para que a biomolécula adote uma conformação mais estável no meio biológico alvo.

O termo conformação é usado para descrever a forma de uma dada molécula, como já empregado neste capítulo. Contudo, deve-se adotar uma distinção entre conformação e estrutura, importante para o entendimento de propriedades moleculares. Estrutura se refere a uma única forma, bem definida e conhecida. Conformação se refere a uma forma dentre múltiplas possíveis, em um determinado meio ou ambiente molecular. Assim, é comum nos referirmos a estrutura cristalina de uma dada proteína, pois no cristal temos uma única forma 3D, como uma foto única que compõe um filme. Em solução, contudo, há diversas formas simultaneamente co-existindo. Neste caso, cada forma pode ser denominada de conformação. Podemos, de forma mais precisa, dizer que a forma de uma biomolécula, determinada por cristalografia de raios-X, é uma conformação cristalográfica.

O processo de enovelamento é mais estudado para proteínas, biopolímeros que apresentam uma versatilidade de estrutura 3ª que nenhuma outra biomolécula possui. Isso faz todo o sentido, tendo em vista que são as proteínas os principais efetores da informação gênica. Em proteínas, o enovelamento envolve a aproximação mútua de resíduos hidrofóbicos, que buscam se esconder da água (também chamado de colapso hidrofóbico), ocasionando a expulsão deste solvente da região central da proteína.

Simultaneamente, os resíduos polares são expostos ao solvente, e interações inter-resíduo são estabelecidas. Assim, a estrutura enovelada, nativa, terá uma quantidade mínima de moléculas de água em seu interior e um número máximo de contatos inter-resíduo (Figura 12-2).

A ideia de ambiente molecular para o enovelamento ou para que uma dada biomolécula exerça sua função é mais complexa do que parece à primeira vista. Embora a ideia usual seja de que o meio aquoso seja predominante, diversos tipos de ambientes aquosos podem ser encontrados dentro de um organismo, tecido ou célula. Por exemplo, o pH pode apresentar grandes variações entre vacúolos lisossomais, citoplasma, plasma, secreção gástrica ou duodenal. Por outro lado, a força iônica da solução pode mudar drasticamente na proximidade de membranas com diferentes cargas.

Outro tipo de ambiente molecular que deve ser destacado é definido pelas membranas biológicas. Membranas são fluidos, e moléculas inseridas em membranas estão solvatados pelas moléculas de fosfolipídeos. Assim, sendo o interior de membranas apolar (ou seja, lipofílico), o colapso hidrofóbico pode acontecer ao inverso, com a exposição de resíduos apolares para o solvente (neste caso, a membrana). Ambientes mais específicos para o enovelamento de proteínas podem ainda ser criados por outras proteínas, denominadas chaperonas. Como um barril, chaperonas podem isolar uma proteína do meio aquoso, levando a formação de interações inter-resíduo que não seriam observáveis de forma significativa em sua ausência. Por conseguinte, podem contribuir diretamente na formação de estruturas 3ªs.

Além de interações não covalentes entre os resíduos de aminoácidos de uma dada proteína (ou as bases de um ácido nucleico e os monossacarídeos de um polissacarídeo) e destes com o solvente, o enovelamento de

proteínas também é influenciado por intera-

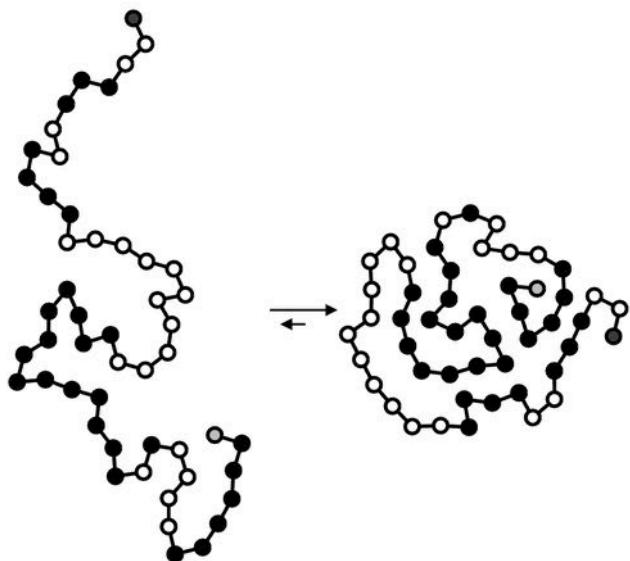



Figura 12-2: Representação 2D do enovelamento de uma proteína hipotética, com o direcionamento de resíduos hidrofóbicos (círculos pretos) para o interior da proteína e dos resíduos hidrofílicos para sua superfície (círculos brancos). Reproduzida de Tomixdf, 2008 (*Creative Commons*).

ções covalentes, associadas a modificações co- ou pós-traducionais.

Durante ou após a síntese proteica (tradução), podem ser formadas ligações dissulfeto entre grupamentos sulfidrila (SH) de resíduos de cisteína, cofatores como o grupamento heme podem ser adicionados ou mesmo processos reversíveis podem ocorrer, nos quais reações como N-acetilação ou fosforilação podem ser observadas de forma transiente. Mas o tipo mais abundante de modificação co- ou pós-traducional na natureza é a glicosilação de proteínas, ou seja, a adição de uma estrutura oligossacarídica a um determinado aminoácido. Assim, a adição destas ligações covalentes e grupamentos altera não somente a forma 3D da proteína, mas sua flexibilidade e múltiplas propriedades fisico-químicas, enzimáticas e, por fim, pode também exercer papel importante em suas funções biológicas.

A glicosilação de proteínas ocorre em mais de 70% das proteínas de eucariotos. Diversos aminoácidos podem estar envolvidos na ligação a carboidratos, mais comumente resíduos de asparagina ou serina, embora também possam participar resíduos de treonina, hidroxiprolina, tirosina, arginina, triptofano e cisteína. Dependendo do aminoácido, a parte sacarídica pode estar ligada a átomos de nitrogênio, oxigênio, carbono ou enxofre, dando origem às glicosilações chamadas de N-, O-, P-, C- ou S-ligadas.

Estrutura 4^{ária}

A despeito da função de um gene ser exercida por uma proteína com estrutura 3D, envolvendo a transmissão de informação de uma estrutura 1^{ária} para uma estrutura 3^{ária}, ainda há um quarto e último nível de organização de biomacromoléculas, denominado de estrutura 4^{ária}. Nem todas as biomoléculas, contudo, apresentam este grau de organização.

A estrutura 4^{ária} é constituída por agregados macromoleculares, principalmente de proteínas. Estas biomoléculas podem adotar estados oligoméricos, sejam estes compostos por 2 (dímeros), 3 (trímeros), 4 (tetrâmeros), 5 (pentâmeros), 6 (hexâmeros) ou mais subunidades necessárias à realização de determinada função em condições nativas. No caso de ácidos nucleicos, a estrutura 4^{ária} também pode ser observada, por exemplo, em complexos entre DNA e proteínas, como histonas.

Não é porque uma proteína se mostra como um oligômero em ambiente cristalino que em solução a mesma organização, necessariamente, será observada. Mesmo *in vivo*, diferentes ambientes fisiológicos podem acarretar em mudanças no estado oligomérico de uma proteína. Por exemplo, um peptídeo que se mostra como monômero no plasma pode formar tetrâmeros quando inserido em membranas.

Portanto, assim como no caso da estrutura 3^{ária}, a estrutura 4^{ária} frequentemente se constitui em uma complexa combinação de múltiplas possibilidades que podem ser modificadas ou reguladas em função de inúmeras variáveis químicas e biológicas. Reproduzir com precisão este comportamento dinâmico é um dos principais desafios para a bioinformática.

2.4. Descritores de forma

O uso dos conceitos de níveis hierár-

quicos nos permite entender as organizações básicas da estrutura 3D de macromoléculas. Estes níveis, contudo, nos oferecem definições qualitativas, gerais, que não abordam nuances ou variações dentro dos níveis. Por exemplo, definir uma região da proteína como uma hélice α não nos informa se esta hélice apresenta ou não algum grau de deformação. Similarmente, podemos saber que uma determinada sequência de nucleotídeos de DNA assume uma hélice do tipo B, mas esta classificação simplemsente não avalia a deformação provocada nesta hélice por um fármaco intercalador do DNA.

Portanto, em acréscimo aos níveis hierárquicos de classificação da estrutura de macromoléculas, há a necessidade de introduzir medidas quantitativas da forma destes compostos. Podemos, assim, calcular precisamente formas associadas a determinados eventos biológicos (como a regulação da expressão de um gene) e, por conseguinte, interferir nestes processos de forma racional (como no desenho de novos fármacos capazes de inibirem a expressão deste gene).

Considerando que proteínas, carboidratos e ácidos nucleicos são biopolímeros, suas formas tridimensionais são definidas, basicamente, pelas conectividades entre seus monômeros constituintes (isto é, aminoácidos, monossacarídeos e bases nitrogenadas, respectivamente).

Esta forma de compreender a estrutura de biomacromoléculas foi proposta inicialmente em 1963 por Gopalasamudram Narayan Ramachandran. Neste trabalho, G. N. Ramachandran descreve a forma de dois aminoácidos vizinhos como fruto dos ângulos de torção ao redor do Cα (Figura 13-2), denominados φ e ψ. Assim, em função das cadeias laterais de cada aminoácido, algumas combinações de ângulos φ e ψ seriam favorecidas, enquanto outras proibidas. As combinações favorecidas correspondem às estruturas 2árias de proteínas que nós conhecemos e oferecem, assim, uma medida quantitativa para definir hélices, fitas, alças e voltas. O gráfico que combina os valores de ângulos φ e ψ para um determinado dipeptídeo ficou assim sendo conhecido como mapa de Ramachandran (Figura 13-2).

O uso de ângulos de torção para descrever a estrutura e a conformação molecular não se limita somente a proteínas, mas também pode ser aplicado a ácidos nucleicos e carboidratos. Em cada caso, o número de ângulos de torção é definido pelas características das ligações entre os monômeros, isto é, se é uma ligação peptídica, glicosídica ou fosfodiéster.

Para a descrição da forma de uma ligação peptídica em uma proteína são empregados três ângulos: ω, ψ e φ. Os ângulos ψ e φ são aqueles descritos no mapa de Ramachandran, localizando-se antes e depois do Cα (porções N- e C- terminais da ligação, respectivamente). O ângulo ω, por sua vez, corresponde ao grupamento amida, ou seja, a ligação entre os grupamentos N-H e C=O (Figura 14-2).

A ligação glicosídica pode ser descrita por dois ou três ângulos torcionais. Em analogia à ligação peptídica, podem ser empregados os ângulos φ e ψ (porção não-redutora e porção redutora, respectivamente). A exceção é quando descrevem-se ligações envolvendo o átomo de carbono na posição 6 de piranoses (como glicose, manose, fucose e etc.) e na posição 5 de furanoses (como na ribose e na desoxirribose). Nestes casos, há a necessidade de se considerar um terceiro ângulo torsional, denominado ω.

O terceiro caso de biopolímeros usualmente descritos por ângulos torcionais, os ácidos nucleicos, consistem em um caso à parte. Como podemos observar na Figura 14-2, o grupamento fosfato agrega grande flexibilidade à cadeia, exigindo assim sete ângulos torsionais para sua adequada caracterização, a saber: α, β, γ (na região 5'), δ (entre os átomos 3' e 4' da pentose), ε e ζ (na porção 3'). Há, ainda, o ângulo χ, formado entre o carbono 1' da pentose e a base nitrogenada.

Ângulos torsionais não são, contudo, a única forma de descrever e avaliar a forma de biomacromoléculas. A despeito de serem biopolímeros, proteínas, carboidratos e ácidos nucleicos apresentam suas particularidades, exigindo assim descritores específicos, capazes de lidar com as propriedades fisico-químicas particulares de cada tipo de monômero (e, por conseguinte, em lidar com as diferentes propriedades biológicas resultantes).

Como mencionado anteriormente, biomoléculas em condições biológicas apresentam não somente uma, mas múltiplas conformações que coexistem, simulta-

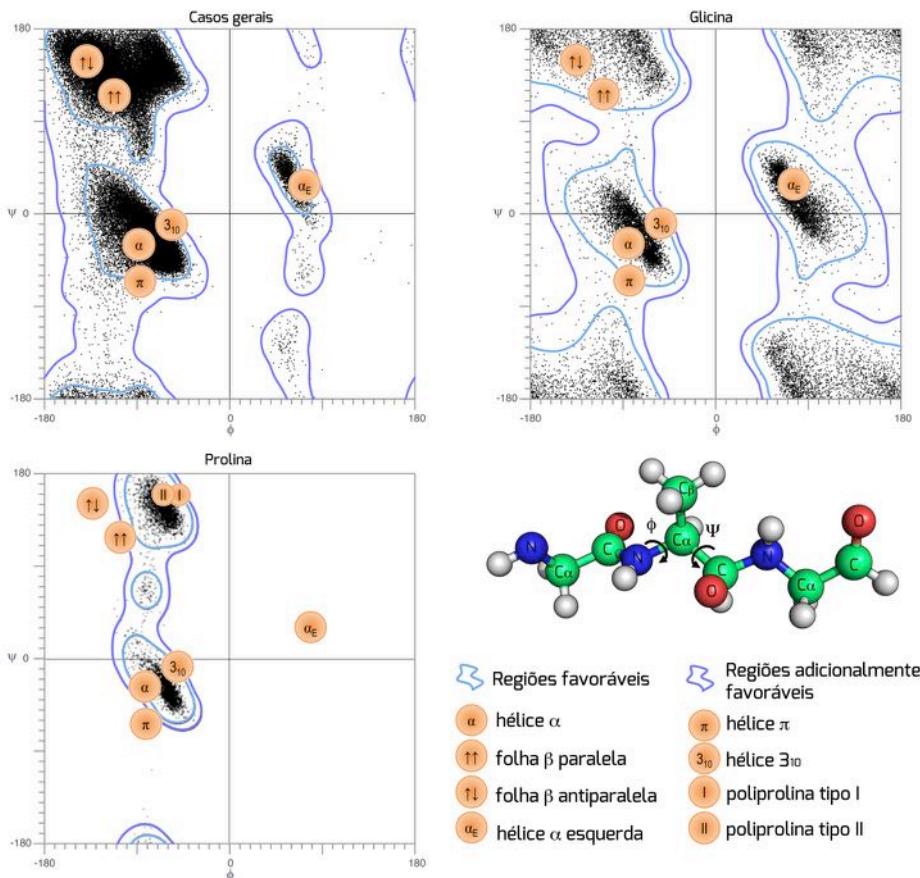



Figura 13-2: Mapas de Ramachandran para casos gerais (resíduos que não sejam prolina ou glicina), para resíduos de glicina e para resíduos de prolina. Os pontos correspondem às distribuições de ângulos φ e ψ de cerca de 100 mil resíduos componentes de 500 estruturas proteicas obtidas em alta resolução. As regiões onde se localizam as estruturas secundárias típicas estão destacadas nos mapas. [Figura baseada em LOVELL, Simon C. *et al*. Structure Validation by Cα Geometry: φ, ψ and Cβ Deviation. Proteins, 50, 437-450, 2003; e Hollingsworth, Scott A. & Karplus, P. Andrew. *A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. Biomol. Concepts*, 1, 271–283, 2010].

neamente. Assim, os valores de ângulos torsionais devem ser considerados como médias, referências geométricas em torno das quais o comportamento da molécula em questão irá variar em solução.

Ácidos nucleicos

Em acréscimo aos ângulos torcionais os ácidos nucleicos, ao formarem pares de bases, definem quase duas dezenas de parâmetros geométricos distintos, importantes para uma cartacterização precisa da estrutura destas biomoléculas (Figura 15-2). Isto ocorre em decorrência de movimentos de translação ou rotação que cada base ou par de bases pode sofrer dentro da região pareada. Assim, moléculas ou regiões de ácidos nucleicos não pareadas não são descritas por estes parâmetros.

Considerando um espaço cartesiano definido pelos eixos x, y e z, sendo z o eixo maior da região de pareamento e bases (Figura 15-2), os parâmetros geométricos oriundos da translação de bases em uma dupla fita envolvem: *i*) o deslocamento do par de bases ao longo do eixo x ou do eixo y; *ii*) o deslocamento de uma base em relação à outra, seja como uma distensão ao longo do eixo y (do inglês *stretch*), seja como cisalhamento ao longo do eixo x (do inglês *shear*), ou ainda um escalonamento acima ou abaixo do plano xy (do inglês *stagger*); *iii*) o deslocamento de um par de base em relação a outro par de base, seja como uma elevação ao longo do eixo z (do inglês *rise*), seja como um deslizamento ao longo do eixo y (do inglês *slide*) ou ao longo do eixo x (chamada em inglês de *shift*).

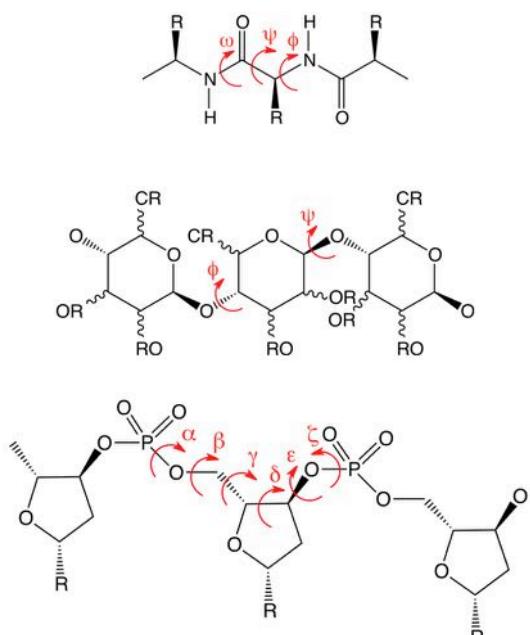



Figura 14-2: Ângulos torsionais para proteínas, carboidratos e ácidos nucleicos ilustrados para, respectivamente, um tripeptídeo, um trissacarídeo e um trinucleotídeo.

Os parâmetros originados da rotação de bases ou pares de bases entre si produzem diferentes tipos de inclinação (definidas em inglês como *tip*, *inclination*, *roll* e *tilt*), dependendo do vértice e do eixo ao longo dos quais ocorre o movimento do par de bases. Pares de bases podem ainda sofrer modificações caracterizando-os como: *i*) torcidos (chamadas em inglês de *twist*, *propeller twist* ou *buckle*), e *ii*) abertos (definida em inglês como *opening*).

Proteínas

Considerando os 20 aminoácidos codificados no genoma, poderíamos imaginar que teríamos 20^n possíveis proteínas diferentes, sendo n o número de aminoácidos. A situação, felizmente, não é tão complexa por uma série de motivos.

Um primeiro aspecto a ser observado é que, quando uma sequência de aminoácidos se enovela para adotar uma determinada estrutura 3ária, alguns aminoácidos se localizam em pontos chave para a estabilização da estrutura 3D. Assim, sua modificação poderia desestabilizar total ou parcialmente a conformação nativa da proteína. Como consequência, algumas posições na sequência de aminoácidos tornam-se conservadas evolutivamente como decorrência de determinantes estruturais. Ao mesmo tempo, podem haver determinantes funcionais para a conservação de posições na sequência ao longo da evolução.

Em contrapartida, como os aminoácidos podem ser agrupados de acordo com a semelhança em suas propriedades fisico-químicas, diferentes combinações de resíduos podem levar a uma mesma estrutura 3D. De fato, sabe-se que a estrutura 3ária de proteínas é mais conservada ao longo da evolução que a estrutura 1ária. Em outras palavras, proteínas com identidade muito baixa entre suas sequências podem possuir estruturas 3árias muito semelhantes.

Conclui-se, assim, que sequências de aminoácidos podem arranjar-se em um conjunto de formas 3D mais ou menos definidos e finitos. Estas formas são denominadas motivos (ou no inglês *fold*), e possuem diversas classificações a partir de suas características (Figura 16-2). Dada a relação entre forma e função, o conhecimento do motivo de uma dada proteína (diretamente por métodos experimentais como cristalografia de raios-X, ver capítulo 13, ou por inferência a partir de similaridade de sequência, ver capítulo 3) é um passo importante para a elucidação de seu mecanismo de ação em nível molecular.

Por exemplo, um barril-β é um motivo que se assemelha a um barri, onde as tiras de madeira correspondem a fitas β (Figura 16-2). Define, assim, uma cavidade central que pode tanto servir como carreador de substâncias, como no caso das nitroforinas, ou como poro, como no caso das porinas. Embora o número de fitas β possa mudar (8 no caso das nitroforinas e 16 no caso das porinas), a característica geral do motivo se mantém. Essas relações são ilustradas visualmente de forma muito elegante na "tabela periódica" de proteínas, desenvolvida pelos professores Richard Garratt e Christine Orengo. Para acessar as classificações dos diferentes motivos já identificados, os bancos de dados CATH e SCOP são as fontes mais completas

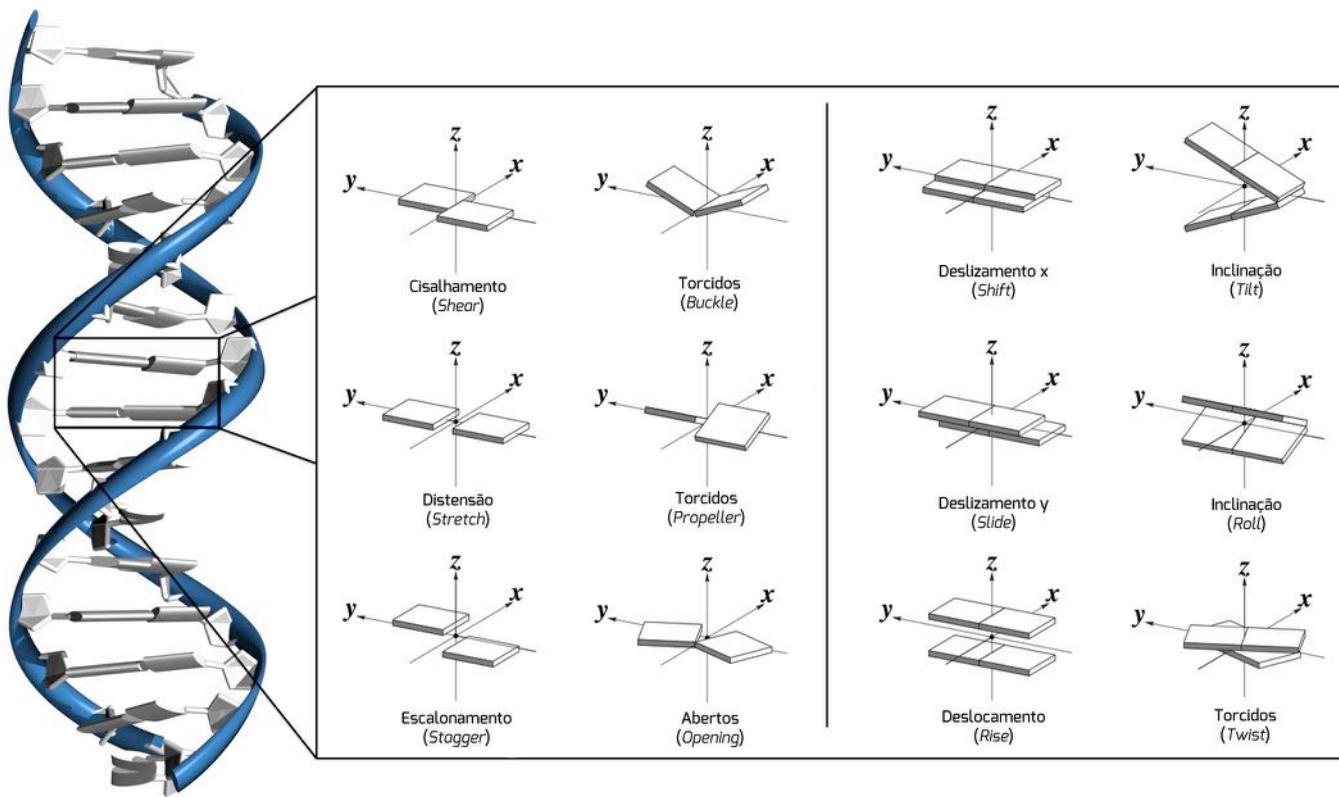



Figura 15-2: Parâmetros geométricos empregados como descritores da geometria de ácidos nucleicos.

de informações.

Um outro conceito, que se confunde e em vários momentos é usado como sinônimo de motivo, é o de domínio proteico. Um domínio é uma parte da sequência polipeptídica de enovelamento independente (e, potencialmente, de função também independente). Assim, se um domínio for recortado de um gene e expresso separadamente ele deve, em princípio, manter suas características estruturais.

Um domínio proteico pode ser composto por mais de um motivo intrinsecamente associado. Por outro lado, um mesmo motivo pode ser encontrado e mais de um domínio de uma mesma proteína.

Membranas

Não temos falado muito de membranas até este momento por alguns motivos. Primeiramente, membranas não são biopolímeros, mas agregados de múltiplas moléculas, o que tira de cena a ideia de análise de uma molécula a partir de suas sub-unidades formadoras. Segundo, estes agregados apresentam-se como um fluido, diferentemente das outras biomoléculas que vimos. Assim, não faz sentido analisar cada molécula de lipídeo individualmente em uma membrana, mas o seu comportamento como um todo ou como uma média ao longo de múltiplos lipídeos.

Contudo, a despeito da natureza fluida de membranas e da sua capacidade de adotar múltiplas formas, os lipídeos (e também proteínas) não se distribuem homogeneamente ao longo das membranas, podendo formar regiões ou domínios enriquecidos em um determinado componente. Assim, para o estudo das propriedades de membranas biológicas torna-se necessário caracterizá-las estruturalmente. Isto pode ser feito através de diversas medidas, tais como a área por lipídeo, espessura da membrana e coeficientes de difusão lateral de lipídeos ou proteínas embebidas na membrana, dentre outros (Figura

8-2).

A área por lipídeo nos oferece informações acerca do grau de compactação das moléculas que constituem uma membrana, ou seja, uma área menor indica uma membrana mais compacta. Isto, por sua vez, sugere uma interação mais intensa entre os componentes da membrana.

Embora proteínas inseridas em membranas adaptem-se a este meio, são as membranas que fazem a maior parte do ajuste em sua estrutura para receber as proteínas (esse processo está relacionado às diferenças de compressibilidade entre estas biomoléculas). Como consequência, a inserção de proteínas em membranas biológicas promove uma perturbação na organização da bicamada lipídica, podendo tanto aumentar quanto reduzir a espessura desta na região ao redor da

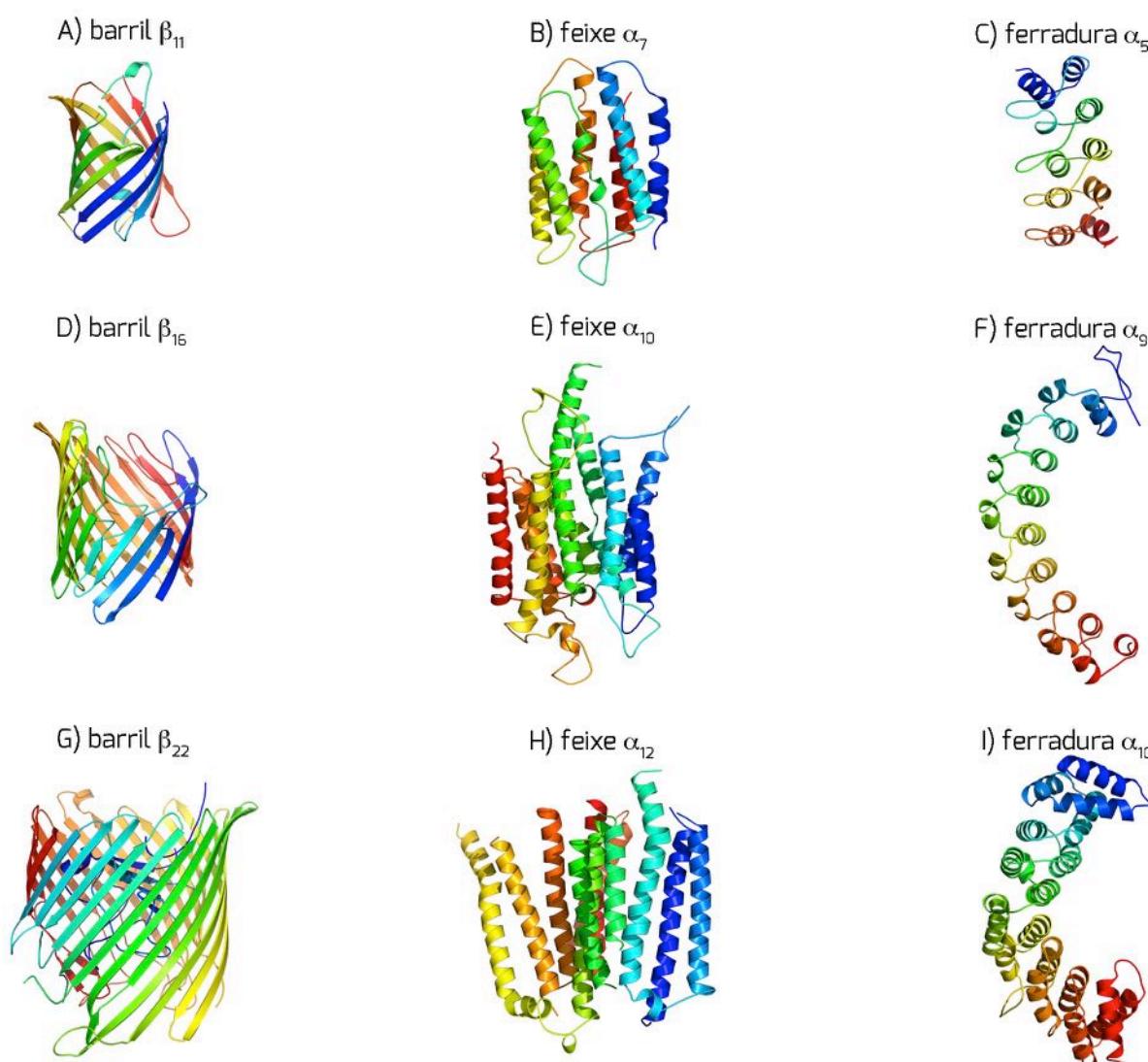



Figura 16-2: Exemplos de motivos proteicos, coloridos por cada elemento de estrutura $2^{\text{ária}}$. São apresentados barris compostos por fitas-β, em A a proteína verde fluorescente (do inglês *green fluorescent protein*, GFP, código PDB 1EMG), em D a porina OMP32 (código PDB 2FGQ) e em G o transportador FECA (código PDB 1KMO); feixes de hélices α, em B a bacteriorodopsina (código PDB 1AP9), em E a proteína SERCA1 (código PDB 1WPG) e em H parte do sistema fotossintético de uma cianobactéria (código PDB 1JB0); e ferraduras compostas por hélices α, em C um inibidor de crescimento tumoral (código PDB 1BD8), em F uma repetição rica em resíduos de leucina, associada à fixação de nitrogênio (código PDB 1LRV) e em H a lipovitelina (código PDB 1LSH). Partes das estruturas foram omitidas buscando maior clareza na imagem. Imagem construída usando o programa Pymol, a partir de organização proposta em "The Protein Chart", de Richard C. Garratt e Christine A. Orengo, 2008, Wiley-VCH.

proteína.

2.5. Formas de visualização

O corolário *uma imagem fala mais do que mil palavras* também se aplica ao estudo de moléculas. E, de fato, o desafio de representar graficamente proteínas vem acompanhando os pesquisadores desde o início dos estudos da estrutura destas moléculas. Os primeiros relatos do uso de representações em cartoon para proteínas datam da década de 1960. Atualmente, múltiplas representações estão à nossa disposição, com qualidade gráfica a cada momento superior, e gerados através de ferramentas gratuitas (Figura 17-2).

Podemos definir hélices de proteínas por suas características geométricas, nomes ou pelos pares de ângulos φ e ψ. Mas visualizar uma hélice proteica, tridimensionalmente, não deixa dúvidas quanto ao seu significado. Portanto, o cuidado com a maneira pela qual iremos apresentar, visualmente, os aspectos estruturais que estudamos e tenhamos relacionados a alguma função biológica, é uma parte fundamental no trabalho do bioinformata.

Formas de visualização, contudo, são representações muitas vezes incapazes de descreverem detalhes sobre a molécula em estudo. É difícil distinguir visualmente uma hélice α de uma hélice 3_{10} ou de uma hélice π. Por outro lado, estas hélices podem apresentar deformações importantes, também de difícil visualização. Assim, a combinação de representações visuais, qualitativas, com medidas precisas, quantitativas, da estrutura molecular é uma estratégia bastante útil no estudo de macromoléculas.

A ideia de combinar múltiplas estratégias na apresentação de um determinado aspecto molecular não se limita somente às formas de descrever visualmente ou numericamente a estrutura molecular. Embora a visualização de estruturas 1^{árias}, isto é, de sequências de nucleotídeos, aminoácidos ou monossacarídeos não nos ofereça muitos artifícios visuais, devemos nos lembrar que as formas apresentadas na Figura 17-2 não informam o leitor facilmente sobre quais resíduos compõe a nossa macromolécula. É difícil distinguir, em representações de arames, bastões ou esferas, uma Ile de uma Leu, e mesmo impossível em cartoon ou superfície. Portanto, pode ser muito útil combinar estas representações tridimensionais a alinhamentos de sequências da região de interesse.

O mesmo vale para a apresentação de sequências isoladas de estruturas. Enquanto uma mutação em um único nucleotídeo pode interferir na função proteica, isso não é feito pela troca de uma letra por outra na sequência, mas por mudanças que esta troca acarretam na estrutura da proteína. O entendimento deste processo pode depender simplesmente da nossa imaginação ou da visualização da respectiva mudança na proteína.

Existem diversas formas de apresentar estruturas tridimensionais de macromoléculas, e escolher entre estas formas envolve tanto escolhas metodológicas quanto pessoais. Algumas propriedades são mais facilmente observadas em alguns tipos de visualização. Por exemplo, o volume da cadeia lateral de um resíduo de Val é muito mais facilmente observável enquanto seus átomos são apresentados como esferas do que como bastões ou arames (Figura 17-2). Diferentes tipos de moléculas, similarmente, se beneficiam de algumas formas de visualização. Por exemplo, a forma de cartoon é a mais comum para descrever proteínas, mas é pouco útil na

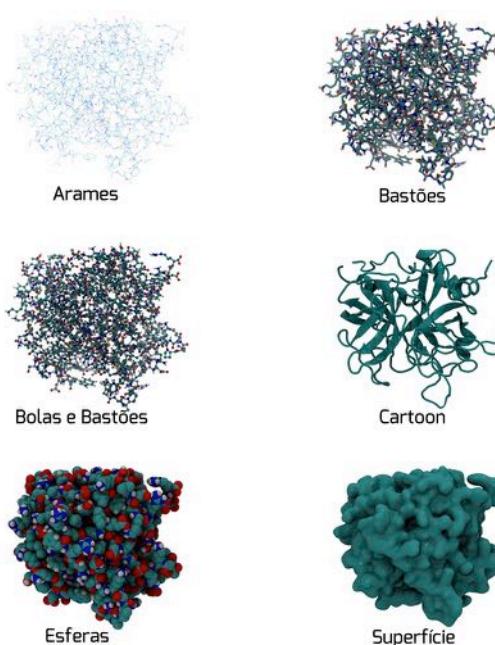



Figura 17-2: Exemplo das formas de visualização mais comumente empregadas na descrição de biomoléculas, aplicadas a uma proteína.

descrição de carboidratos ou membranas.

Em muitos casos poderemos empregar combinações destas formas, como na descrição por cartoon de uma proteína e de sua estrutura de glicosilação como bastões.

2.6. Conceitos-chave

Anfipatia: propriedade de moléculas que possuem tanto regiões hidrofílicas quanto hidrofóbicas.

Cadeia lateral: região variável dos aminoácidos codificados no genoma, responsável pela variação de suas propriedades.

Carbono anomérico: átomo de carbono numerado como 1 em carboidratos. A mudança em sua estereoquímica dá origem às formas anoméricas α e β em carboidratos.

Carbono α: átomo de carbono do esqueleto peptídico no qual a cadeia lateral de cada aminoácido está ligada (referindo-se aos 20 aminoácidos codificados no genoma para síntese proteica). É o primeiro átomo de carbono vizinho ao grupo carbonila.

Conformação em bote torcido: forma adotada pelo anel de alguns monossacarídeos.

Conformação em cadeira: forma adotada pelo anel de alguns monossacarídeos, semelhante a uma cadeira quanto vista de lado.

Conformação em envelope: forma adotada pelo anel de alguns monossacarídeos, destacadamente as furanoses.

Dogma central da biologia molecular: representação do fluxo de informação em sistemas biológicos, começando na molécula de DNA e culminando na síntese proteica - mas não no sentido oposto. Envolve principalmente os fenômenos de replicação, transcrição e tradução.

Enovelamento: processo segundo o qual uma sequência polipeptídica adquire sua estrutura tridimensional nativa, isto é, equivalente àquela observada em seu local biológico de ação e funcional. Também chamado por alguns autores de dobramento.

Equilíbrio pseudo-rotacional: processo de interconversão entre as diferentes conformações adotadas por carboidratos.

Esqueleto do DNA: parte da molécula de DNA composta pelas partes comuns a todos os nucleotídeos, isto é, o carboidrato e o grupo fosfato (ou seja, são excluídas as regiões das bases nitrogenadas).

Esqueleto peptídico: estrutura de peptídeos ou proteínas sem as cadeias laterais dos aminoácidos (ou seja, somente as regiões comuns aos aminoácidos).

Estrutura $1^{ária}$: sequência de letras que compõe biomoléculas (principalmente DNA, RNA e proteínas, mas também carboidratos).

Estrutura $2^{ária}$: padrões estruturais definidos pela organização das unidades monoméricas (isto é, nucleotídeos, aminoácidos e monossacarídeos) de cada biomolécula em formas tridimensionais. Estes padrões podem classificados segundo suas diferentes formas.

Estrutura $3^{ária}$: estrutura 3D completamente enovelada.

Estrutura $4^{ária}$: organização definida pela agregação de múltiplas estruturas $3^{árias}$.

Furanoses: monossacarídeos cujo anel é composto por 5 átomos, quatro de carbono e um de oxigênio. O nome vem da semelhança deste anel com o composto furano.

Ligação fosfodiéster: ligação formada entre dois nucleotídeos, através de seus grupos fosfato.

Ligação glicosídica: ligação formada entre dois

monossacarídeos.

Ligação peptídica: ligação formada entre dois aminoácidos, através do grupo amino de um resíduo e do grupo carboxila do outro, dando origem a uma função amida.

Mapa de Ramachandran: um gráfico que descreve a variação da energia em função da rotação dos ângulos de diedro φ e ψ, ao redor do $C\alpha$.

Nucleosídeo: molécula formada por uma base nitrogenada ligada a um carboidrato (ribose ou desoxirribose), sem o grupo fosfato.

Nucleotídeo: molécula formada por uma base nitrogenada ligada a um carboidrato (ribose ou desoxirribose) e a um grupo fosfato.

Piranoses: monossacarídeos cujo anel é composto por 6 átomos, cinco de carbono e um de oxigênio. O nome vem da semelhança deste anel com o composto pirano.

2.7. Leitura recomendada

ALBERTS, Bruce; et al. ***Biologia Molecular da Célula***. 5.ed. Porto Alegre: Artmed, 2010.

BLOOMFIELD, Victor A.; CROTHERS, Donald M.; TINOCO, JR., Ignacio. ***Nucleic Acids Structure, Properties, and Functions***. Sausalito: University Science Books, 2000.

GARRATT, Richard C., ORENGO, Christine A. ***The Protein Chart***. Nova Iorque: Wiley-VCH, 2008.

PETSKO, Gregory A.; RINGE, D. ***Protein Structure and Function***. New York: Oxford University Press, 2009.

3. Alinhamentos

Dennis Maletich Junqueira
Rodrigo Ligabue Braun
Hugo Verli

```
      ATATACTTGATCGATCAGCA              ATATACTTGATCGATCAGCA
ATATACTTG---TTCCATCAGCT                 ATATACTTGTTCCATCAGCT
   ATATA----CTTGTTCCATCAGCT      →      ATATACTTGTTCCATCAGCT
AT---ATACT----TGATCGA---TTAGCT          ATATACTTGATCGATTAGCT
                                        *********  **  **  ***
```

Alinhamento de múltiplas sequências.

3.1. Introdução

O avanço nas técnicas de sequenciamento do DNA tem permitido um crescente aumento no número de genomas disponíveis em bancos de dados públicos. Esta maior disponibilidade exigiu um grande aumento na capacidade computacional de armazenamento e no investimento em desenvolvimento de técnicas de processamento adequadas para a análise destes dados. Algoritmos de análise tiveram de ser criados e aperfeiçoados e, dentre estes, as técnicas de alinhamento de sequências tornaram-se ferramentas essenciais e primordiais na análise de sequências biológicas. Atualmente, diversos programas *online*, ou mesmo de instalação local, são capazes de alinhar centenas de sequências em poucos minutos.

Devido à extensão de suas aplicações, o alinhamento de sequências biológicas é um processo de fundamental importância para a bioinformática. Conceitualmente, os alinhamentos são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas.

Em geral, as moléculas consideradas por estes programas, sejam elas formadas por nucleotídeos (DNA ou RNA) ou aminoácidos (peptídeos e proteínas), são polímeros representados por uma série de caracteres, e a comparação entre as moléculas depende apenas da comparação entre as respectivas letras. Apesar da facilidade e da aparente simplicidade do processo, a análise de similaridade das sequências é uma tarefa complexa e uma etapa decisiva para grande parte dos métodos de bioinformática que fazem uso de sequências biológicas.

Durante o alinhamento, as sequências são organizadas em linhas e os caracteres biológicos integram as colunas do alinhamento (Figura 1-3). Seguido à organização inicial, algoritmos específicos buscarão a melhor correspondência para as sequências em questão, permitindo a criação de espaços entre estes caracteres para que, ao final, todas as sequências tenham o mesmo comprimento. Isto possibilita uma fácil visualização da similaridade, permitindo que caracteres

Figura 1-3: Alinhamento de quatro sequências de nucleotídeos envolvendo 55 caracteres. *a*) Grupo de sequências não alinhadas, cada sequência ocupando uma linha individual. *b*) Grupo de sequências alinhadas, onde caracteres idênticos são dispostos em uma mesma coluna e estas são identificadas por asteriscos (dispostos na parte inferior do alinhamento). Nucleotídeos ausentes em determinadas sequências são substituídos por hifens para identificar eventos de inserção/deleção.

idênticos ou similares em cada uma das sequências integrem a mesma coluna. A ideia central destes algoritmos é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. Comumente, a similaridade entre as sequências envolvidas é expressa pelo termo identidade, que quantifica a porcentagem de caracteres idênticos entre duas sequências.

A relevância e abrangência do uso do método tornam os procedimentos de alinhamento o cerne para diferentes campos dentro da grande área da bioinformática. Além de fundamentais em pesquisas de filogenética e análise evolutiva, os alinhamentos são exigidos em estudos de inferência estrutural e funcional de proteínas, análises de similaridade e identificação de sequências e em estudos aplicados ao campo da genômica.

Através dos métodos de alinhamento, é possível obter informações a respeito da relação evolutiva entre organismos, indivíduos, genes ou entre sequências diversas (Figura 2a-3). Se duas sequências distintas podem ser alinhadas com certo grau de similaridade, é possível inicialmente assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, por isso, são evolutivamente relacionadas. A partir da separação destas sequências de seu ancestral comum, individualmente cada uma delas acumulou diferentes variações ao longo do processo evolutivo. O termo homologia é utilizado frequentemente para definir estes eventos onde, através da relação de ancestralidade, dois indivíduos distintos possuem regiões em seu DNA (incluindo regiões codificantes) herdadas de um ancestral comum. Neste caso, a similaridade deve-se à descendência comum e, portanto, as sequências envolvidas na análise são ditas homólogas.

Cabe ressaltar que a homologia não requer necessariamente alta identidade de caracteres entre as sequências, uma vez que a maior ou menor identidade entre elas dependerá da taxa de evolução do organismo ou da espécie (consultar capítulo 5). Ainda, a similaridade entre sequências pode ser gerada não somente por descendência, mas por pressão seletiva de um determinado ambiente. Nestes casos, teremos regiões similares na sequência de nucleotídeos (ou aminoácidos) que surgiram de maneira independente, sem qualquer relação de descendência, e evoluíram por convergência, não sendo portanto homólogas. Assim, não é possível quantificar a homologia entre as sequências envolvidas, somente dizer se há ou não. Quando identificamos quantos caracteres se repetem nas mesmas posições entre duas ou mais sequências estamos, de fato, verificando a identidade entre estas, e não a homologia.

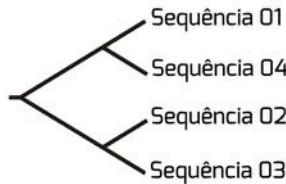
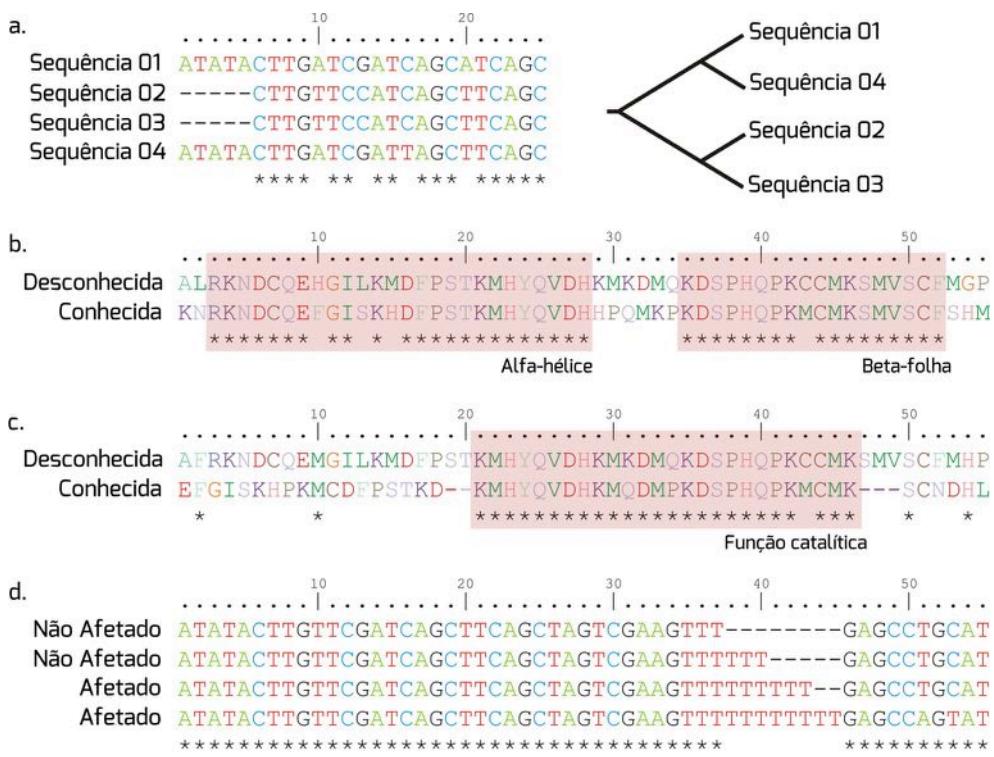



Figura 2-3: Aplicações dos métodos de alinhamento de sequências biológicas. *a*) Inferência filogenética a partir do alinhamento de quatro sequências de nucleotídeos. *b*) Inferência da estrutura de uma proteína alvo (Desconhecida) a partir do alinhamento com uma sequência de aminoácidos cuja estrutura tridimensional é conhecida (Conhecida). *c*) Inferência da função de um domínio proteico a partir da comparação de sequências de aminoácidos. *d*) Comparação de sequências de uma porção de determinado gene de indivíduos afetados e não afetados por uma doença genética. Os asteriscos identificam colunas com total similaridade dos caracteres.

As técnicas de alinhamento vêm se mostrando fundamentais na construção de algoritmos que visam comparar a informação de diversas sequências biológicas. À exemplo do programa BLAST, estes algoritmos permitem comparar uma sequência alvo com milhares de dados disponíveis em grandes bancos de armazenamento, fornecendo um valor de significância estatística associada a esta comparação de similaridade. Devido à facilidade de acesso e rapidez no processamento de dados, estes programas vêm cada vez mais ampliando as possibilidades e opções para o tipo de comparação ou pesquisa a ser realizada.

Os métodos de alinhamento podem ainda ser necessários para fornecer informações a respeito da função e da estrutura de sequências biológicas, particularmente nos alinhamentos de ribonucleotídeos e aminoácidos (Figura 2-3). Nestes casos, a similaridade entre duas ou mais sequências (dada em porcentagem) revela padrões referentes à composição química e podem fornecer embasamento para a definição de um arranjo tridimensional semelhante, principalmente no caso de proteínas (Figura 2b-3). A mesma relação é feita para inferir a função de domínios de uma proteína recém-descoberta, ainda sem função definida. Sabendo que sua forma está diretamente relacionada à sua função, através da comparação com outras proteínas com estrutura e função já estabelecidas, é possível inferir a função realizada por determinado domínio da proteína sob investigação (Figura 2c-3). Nestes casos, as sequências envolvidas no alinhamento não são necessariamente homólogas. Através do fenômeno da evolução convergente, diferentes regiões codificantes do DNA podem gerar produtos proteicos com funções similares, sem obrigatoriamente compartilharem um ancestral comum.

Finalmente, as técnicas de alinhamento

têm grande importância para a análise de genes e genomas. Com o aumento da disponibilidade de sequências nucleotídicas de genomas completos, e mesmo com o surgimento de modernas técnicas de biologia molecular, como o *microarray* e *deep sequencing*, os métodos de comparação permitiram o entendimento a respeito da variabilidade genética de indivíduos e populações.

A comparação entre genomas de diferentes espécies, ou até mesmo de indivíduos da mesma espécie, possibilita a análise de variações (mutações ou polimorfismos) nas sequências e, em alguns casos, permite a identificação de relações entre variações no DNA e susceptibilidade a determinadas doenças, beneficiando o campo da genética e áreas relacionadas. Adicionalmente, como um recurso para a caracterização de eventos evolutivos, os alinhamentos permitem análises comparativas entre genomas. A abrangência e importância evolutiva dos eventos de quebra e reparo de DNA, ou mesmo dos eventos de recombinação, inversões e translocações, tem sido desvendados, primariamente, através dos métodos de alinhamento.

Além do alinhamento de sequências, o alinhamento de estruturas constitui outra importante ferramenta em estudos de bioinformática. A metodologia é bastante diferente daquela empregada em alinhamentos de sequências, pois passamos de um problema unidimensional para um problema tridimensional. Sua utilização passou a ser difundida a partir de 1978, com o trabalho de Rossmann e Argos, comparando os sítios ativos de enzimas cujas estruturas eram conhecidas até aquele momento. Os métodos de sobreposição simples de estruturas estão disponíveis há mais tempo, tendo sido propostos a partir da década de 1970, enquanto os métodos de comparação e alinhamento se desenvolveram posteriormente, principalmente a partir da década de 1990.

A comparação de estruturas se refere à análise de similaridades e diferenças entre duas ou mais estruturas, enquanto o alinhamento de estruturas se refere à determinação de quais aminoácidos seriam equivalentes entre tais estruturas. É importante destacar também a diferença entre alinhamento e sobreposição de estruturas. Apesar desses termos ainda serem empregados na literatura como sinônimos, eles se referem a procedimentos diferentes. Conforme mencionado acima, enquanto o alinhamento de estruturas busca identificar equivalências entre pares de aminoácidos nas estruturas a serem sobrepostas, a sobreposição necessita desse conhecimento prévio sobre as equivalências.

Sendo assim, a sobreposição estrutural busca solucionar um problema muito mais simples, ou seja, minimizar a distância entre dois resíduos já reconhecidos como equivalentes. Isso se dá por encontrar transformações que satisfazem o menor desvio médio quadrático (RMSD) ou as equivalências máximas dentro de um valor limite para o RMSD.

Considerando que a estrutura das proteínas é mais conservada que a sequência, o alinhamento de estruturas confere maior especificidade ao alinhamento de sequências quando comparado ao alinhamento de sequências independente de estrutura. A maioria dos métodos de sobreposição de estruturas é adequado para identificar similaridades entre estruturas proteicas. O alinhamento de duas ou mais estruturas, porém, constitui uma tarefa mais difícil, e sua precisão depende tanto do método usado quanto do objetivo do usuário.

3.2. Alinhando sequências

À primeira vista, o processo de alinhamento entre diferentes sequências parece simples e não sujeito a qualquer tipo de erro. No entanto, esta afirmativa só é verdadeira em casos onde os organismos envolvidos possuem uma baixa taxa evolutiva (Figura 3a-3). Quando consideramos sequências homólogas amostradas de organismos com alta taxa evolutiva, ou até mesmo sequências similares, porém não homólogas, nos deparamos com casos particulares que tornam o processo de alinhamento complexo e, muitas vezes, sujeito a uma interpretação especialmente subjetiva por parte do usuário (Figura 3b-3).

A comparação de sequências homólogas de organismos evolutivamente distantes é um desafio para os programas de alinhamento. As diferentes pressões seletivas moldam os genomas de maneira imprevisível e, muitas vezes, acarretam a perda ou ganho de nucleotídeos ao longo do processo evolutivo. Para estes casos, a adição de lacunas (*gaps*) em matrizes de alinhamento, representadas por "-", é possível e muitas vezes necessária. As lacunas representam um ou mais eventos de inserção ou deleção de nucleotídeos. Estes eventos, comumente chamados de "indels" (*in* para inserção, e *del* para deleção), são fruto de processos mutagênicos (espontâneos ou induzidos) e, dependendo da região atingida, podem ser expressos nas moléculas de RNA e nas proteínas, onde poderão gerar consequências moleculares. Erros de replicação gerados pela DNA-polimerase durante a replicação do DNA, ou mesmo os eventos de recombinação, são os principais fatores atrelados à geração destes *indels* nos genomas. Em regiões codificadoras, estes eventos podem acarretar mudanças no quadro de leitura da proteína e torná-la não funcional.

Em termos analíticos, a inserção de lacunas dificulta o processo de alinhamento e exige interpretações cautelosas. Para determinados casos, especialmente em análises evolutivas e filogeográficas, é comum que regiões do alinhamento com determinado nível de incerteza, especialmente regiões com grande número de lacunas, sejam eliminadas

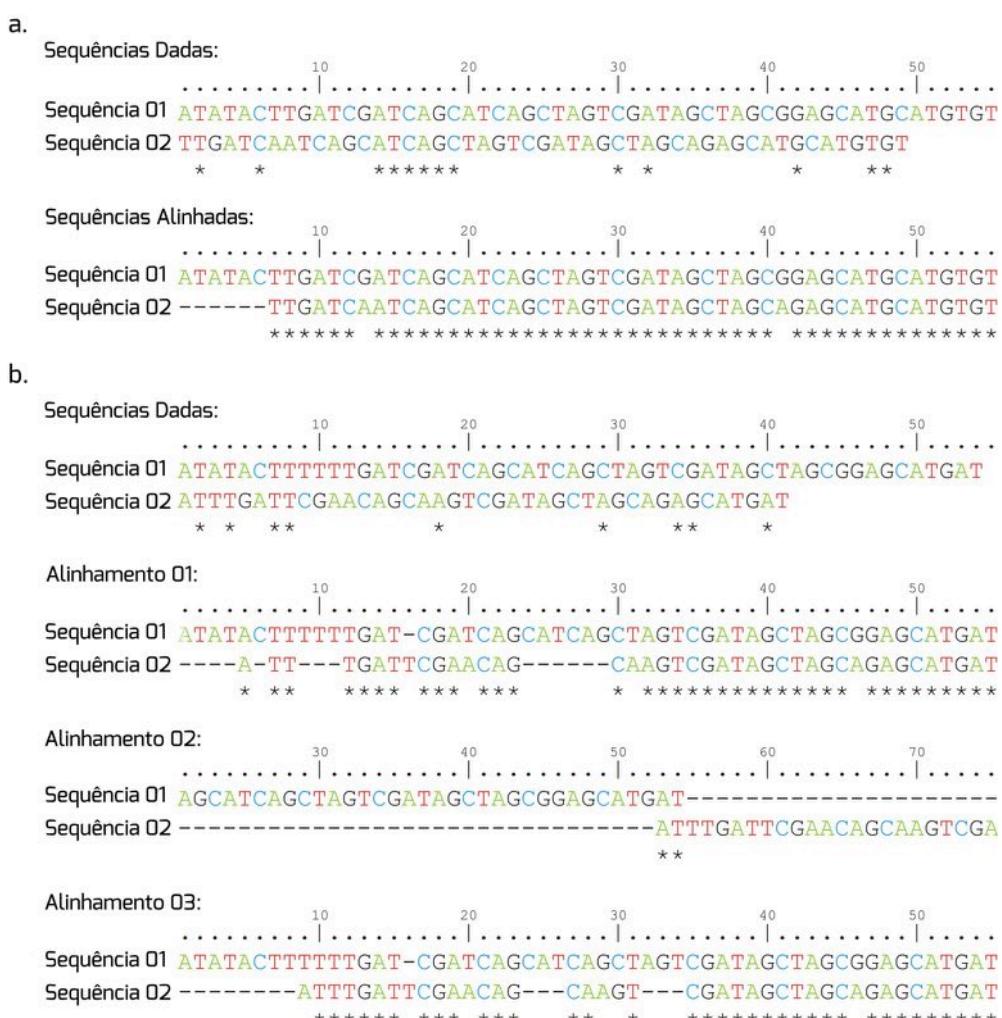



Figura 3-3: Alinhamentos de nucleotídeos. *a*) Duas sequências homólogas originadas de organismos com baixa taxa de evolução são dadas e seu alinhamento é proposto. *b*) Duas sequências homólogas amostradas de organismos com alta taxa de evolução são dadas e diferentes alinhamentos são propostos. Os hifens representam eventos de inserção ou deleção únicos na sequência. Os asteriscos identificam colunas com total similaridade dos caracteres.

da análise. Contudo, até o momento não existem programas capazes de lidar com as lacunas de forma coerentemente biológica. Apesar de sabermos que se tratam de eventos evolutivos comuns e bem caracterizados, as incertezas sobre o número de eventos e sua intensidade tornam as lacunas, em grande parte dos casos, um fator de confusão para análises de alinhamento.

Conforme mostrado na Figura 3-3, diferentes alinhamentos são possíveis para um mesmo grupo de sequências. A pergunta que se segue é: como reconhecer o melhor resultado quando nos deparamos com diversos alinhamentos possíveis para um mesmo conjunto de dados? Buscou-se resolver este problema através da criação de um sistema de pontuação para comparar os resultados de diferentes alinhamentos. Caracteres idênticos em sequências diferentes representam igualdades ou correspondências (*matches*) e, por serem resultados preferenciais durante o processo de alinhamento, são pontuados positivamente. Pelo contrário, caracteres não idênticos que ocupam a mesma coluna são chamados de desigualdades, ou *mismatches*, e recebem atribuições negativas. Como resultado, o melhor alinhamento possível para duas sequências é aquele que maximiza a pontuação total, somando os valores de *matches* e debitando os valores de *mismatches*.

Do ponto de vista biológico, as mudanças entre as bases nitrogenadas nas sequências de nucleotídeos não ocorrem com a mesma probabilidade (Figura 4a-3). Sendo assim, podemos atribuir valores de *mismatches* diferentes às transições (trocas de purinas por purinas ou pirimidinas por pirimidinas) e às transversões (trocas de purinas por pirimidinas ou pirimidinas por purinas). Para sequências de aminoácidos, é necessário escolher ativamente uma matriz de pontuação específica. Essas matrizes são resultados diretos de estudos de variação proteica e estão diretamente relacionadas à probabilidade de substituição de um aminoácido por outro (matrizes BLOSUM e PAM). Atualmente, as matrizes BLOSUM são as mais disseminadas e aplicadas para os mais diversos casos de comparação entre sequências de aminoácidos (Figura 4b-3).

a.

	A	C	G	T
A	1	-2	-2	-2
C		1	-2	-2
G			1	-2
T				1

b.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-2	-2	0	0	0	0	-2	-2	-3	-2	-1	-2	0	0	0	-2	-3	0
R		5	-2	-3	-3	0	-1	-2	0	-3	-4	1	-3	-3	-2	-2	0	0	-3	-4
N			5	0	0	0	-2	0	0	-4	-5	-2	-3	-3	-2	0	0	-2	-2	-5
D				5	-4	0	1	-1	0	-5	-6	-3	-4	-4	0	-2	-2	-2	-2	-5
C					8	-2	-3	-1	-1	0	-2	-3	0	-1	-1	1	0	0	-2	0
Q						5	2	0	0	-2	-4	0	-2	-3	0	0	0	0	-2	-3
E							5	0	0	-3	-4	0	-3	-3	0	0	0	-2	-3	-3
G								6	0	-4	-5	-2	-3	-2	-2	0	0	0	-2	-3
H									6	-3	-4	0	-2	0	0	0	0	0	2	-2
I										4	0	-3	2	0	-2	-3	0	0	-3	2
L											4	-4	0	0	-3	-4	-3	0	-4	0
K												4	-2	-4	-1	-2	0	0	-3	-4
M													6	0	-3	-3	-2	0	-3	2
F														6	-3	-2	-2	2	2	0
P															7	0	0	-2	-3	0
S																4	2	-2	-2	-3
T																	5	-1	-3	0
W																		9	2	-1
Y																			7	-3
V																				4

Figura 4-3: Matrizes de custo utilizadas no cálculo de pontuação dos alinhamentos. *a*) Matriz de custo exemplo utilizada para cálculos de pontuação em alinhamentos de nucleotídeos. *b*) Matriz de custo BLOSUM62 utilizada para cálculo da pontuação em alinhamentos de aminoácidos.

Ainda, é necessário que as lacunas de alinhamentos recebam determinadas pontuações, pois são frequentemente encontradas em alinhamentos de dados biológicos. Se lacunas podem ser adicionadas em qualquer posição sem qualquer restrição, tanto nas extremidades quanto no interior das sequências, é possível gerar alinhamentos com mais lacunas do que propriamente caracteres a serem comparados (Figura 3b-3, alinhamento 2). Com o intuito de prevenir inserção excessiva, a adição de lacunas é penalizada durante a atribuição da pontuação de uma sequência, conforme um conjunto de parâmetros, chamado de penalidades por lacuna (*gap penalties*, PL). A abrangência da lacuna é pontuada pelo respectivo número de *indels* presentes no alinhamento. A fórmula mais comum para cálculo destas penalizações segue abaixo:

$$PL = g + e\,(L - 1)$$

onde L é o tamanho da lacuna (número de *indel*s presentes na lacuna), g é a penalidade pela abertura da lacuna (necessária para evitar que os alinhamentos contenham lacunas desnecessárias) e e é a penalidade atribuída a

cada *indel* (novamente para evitar grandes lacunas sem necessidade). Os valores de penalidade por lacuna são desenhados para reduzir a pontuação de um alinhamento quando este possui uma quantidade de *indels* desnecessária. Apesar da disseminação deste conceito, não há qualquer relação matemática ou biológica sustentando este cálculo. É importante destacar que, através da propriedade de "alinhamento livre de colunas em branco" (ou seja, *gaps* não são alinhados), as penalizações ainda impedem o alinhamento de *indels* entre as sequências envolvidas na análise. Assim, o melhor alinhamento entre as sequências será dado por um valor que resulta da soma dos valores associados a cada um dos *matches*, *mismatches* e lacunas, de acordo com um critério pré-definido (Figura 5-3).

O método de pontuação foi a solução encontrada para avaliar e classificar diferentes alinhamentos em busca da melhor explicação para a relação evolutiva entre as sequências. O próximo problema encontrado foi enumerar todas as possibilidades de alinhamentos para um grupo de dados. Assumindo-se duas sequências com tamanho de 100 caracteres cada, poderíamos enumerar até 10^{77} possíveis alinhamentos, diferentes entre si. A extensão de possibilidades inviabiliza a enumeração de todos os casos devido ao tempo e ao requerimento de enorme processamento destes dados. Apesar da exigência computacional, alguns algoritmos são capazes de realizar tal tarefa e ainda aplicar o método de pontuação para cada um dos casos, em busca do melhor resultado. No entanto, estes algoritmos não são capazes de lidar com sequências que contenham mais que algumas dezenas de caracteres. Em virtude da capacidade de explorar todas as soluções do problema, o processo realizado por estes algoritmos é chamado de "alinhamento ótimo".

Contudo, em virtude da inerente demora do processo, foi necessário desenvolver algoritmos que acelerassem a busca de um alinhamento capaz de explicar de maneira ótima os processos evolutivos para um determinado grupo de sequências sem, no entanto, enumerar todas as possibilidades. Os alinhamentos gerados por estes programas são chamados heurísticos, e compreendem métodos aproximados de busca pelo resultado ótimo. Diferentes métodos foram criados para diferentes tipos de alinhamento (Figura 6-3). Entre estes, devido à eficiência e à rapidez de processamento das informações de um alinhamento, incluindo o cálculo de pontuação, os algoritmos de programação dinâmica são, atualmente, os mais utilizados para este fim, tanto em alinhamentos simples como integrado aos algoritmos de alinhamentos múltiplos.

É fundamental assumirmos, para a maior parte dos problemas em bioinformática, o alinhamento como um modelo de relação evolutiva entre as sequências envolvidas. E como modelo, está sujeito à presença de certos problemas na explicação dos eventos evolutivos reais. Portanto, os alinhamentos devem ser avaliados com extrema cautela. A facilidade e a aparente simplicidade na análise dos programas tornam o processo mecânico e desvinculado de análises críticas pela maior parte dos usuários. A associação dos métodos de alinhamento a outras análises de bioinformática tende a desvincular a real importância desta técnica e a coloca apenas como um procedimento, e não formalmente como uma técnica sujeita à análise crítica. Isto pode ocasionar na obtenção de modelos incorretos ou mesmo de falsos positivos.

3.3. Tipos de alinhamento

Em estudos de bioinformática, é comum compararmos moléculas de dois ou mais indivíduos, sejam eles da mesma espécie ou de espécies diferentes. Quanto maior o número de sequências comparadas, maior o tempo exigido para conclusão do alinhamento e, dependendo das sequências envolvidas, maior a dificuldade dos algoritmos em encontrar o melhor resultado. Conforme a quantidade de sequências envolvidas, podemos dividir os alinhamentos em dois tipos: alinhamentos simples, ou par-a-par, e alinhamentos múltiplos, ou de múltiplas sequências (Figura 7-3).

Figura 5-3: Esquema de pontuação para avaliação de alinhamentos. *a*) Duas sequências de desoxirribonucleotídeos não alinhadas. *b*) Proposição de um alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e duas lacunas com 3 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -7. *c*) Proposição de um segundo alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e três lacunas com 1, 2 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -8. A partir deste exemplo, o alinhamento com a maior pontuação é o mostrado em *b*. Os valores de pontuação utilizados neste exemplo são especificados na parte inferior da figura.

Os alinhamentos simples descrevem especificamente a relação de similaridade entre duas sequências quaisquer. Já os alinhamentos múltiplos incluem três ou mais sequências na análise de similaridade e, dependendo do objetivo do usuário, podem envolver até centenas de sequências.

Conceitualmente, ainda podemos dividir os alinhamentos, tanto simples, como múltiplos, em dois grandes tipos. Os alinhamentos que levam em consideração toda a extensão das sequências são conhecidos como globais, enquanto aqueles que buscam pequenas regiões de similaridade são chamados de locais (Figura 7-3). Em algoritmos que buscam o alinhamento global de duas sequências, reforça-se a busca do alinhamento completo das sequências envolvidas, procurando incluir o maior número de *matches* do início ao final das sequências. Quando necessário, estes algoritmos permitem a inserção de lacunas para que as sequências tenham o mesmo tamanho no resultado do alinhamento (Figura 7b-3).

Graficamente, os sítios com caracteres idênticos são representados ligados por barras verticais, enquanto os sítios que possuem caracteres diferentes nas duas sequências, ou

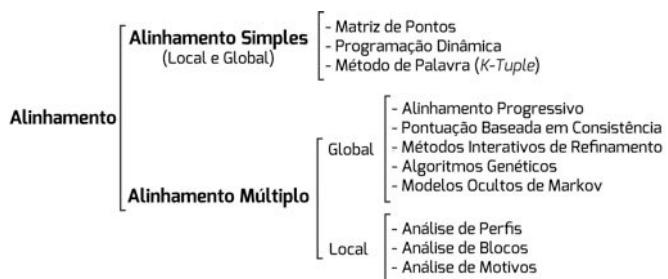



Figura 6-3: Tipos de alinhamento e os algoritmos aplicados à bioinformática.

mesmo a presença de uma lacuna em uma delas, permanecem sem qualquer notação (Figura 7-3). O principal algoritmo envolvido no processamento de alinhamentos globais é aquele desenvolvido por Needleman e Wunsch durante a década de 1970. Além de ter uma notável importância metodológica, este algoritmo tem grande importância na história do alinhamento, pois foi o primeiro algoritmo a aplicar o método de programação dinâmica para a comparação de sequências biológicas.

Em seu início, os métodos de alinhamento eram utilizados especialmente para a comparação par-a-par de sequências de proteínas inteiras. No entanto, com a ampliação da disponibilidade de sequências completas de proteínas, foi necessário buscar métodos de alinhamento que privilegiassem a busca de similaridade, não entre sequências completas, mas apenas entre porções isoladas destas sequências. Durante a década de 1980 iniciou-se o desenvolvimento de novos algoritmos de alinhamento, já que os desenvolvidos até aquele momento não eram aplicáveis para esta particularidade. Entre estes novos algoritmos, o desenvolvido por Smith e Waterman, em 1981, ganhou maior destaque e atualmente é o principal algoritmo utilizado por programas para realização de alinhamentos locais. Nestes casos, privilegia-se o alinhamento de partes da sequência, buscando apenas as regiões com a maior similaridade (Figura 7c-3). Em algoritmos para busca local, o alinhamento pára no final das regiões de alta similaridade e substitui as regiões excluídas por hifens (lacunas) no resultado final (Figura 7c-3).

3.4. Alinhamento simples

Para entender como se processa um alinhamento par-a-par e como o grau de si-

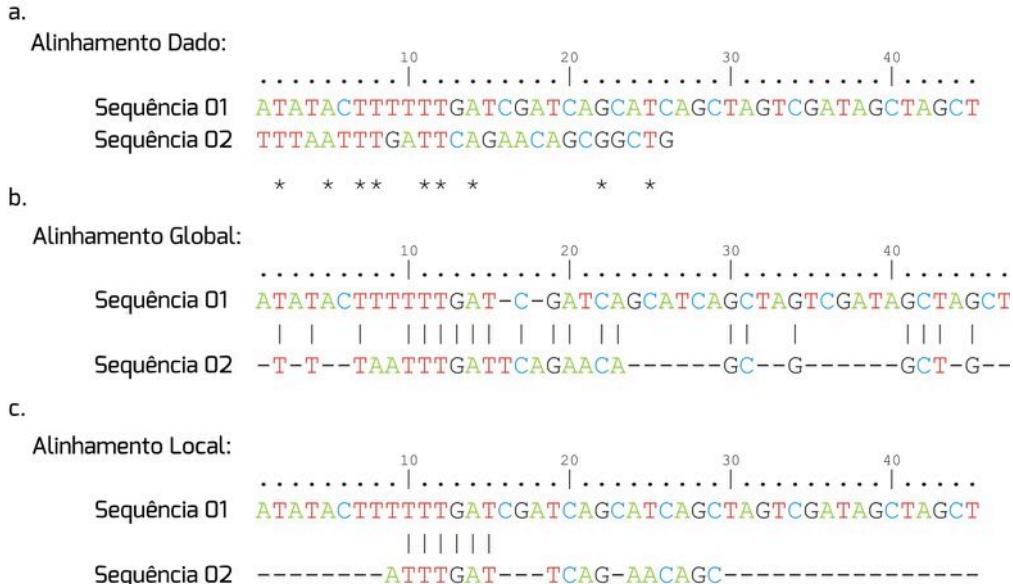



Figura 7-3: Diferenças entre alinhamento local e global. *a*) Duas sequências de nucleotídeos de tamanhos diversos são amostradas e alinhadas por algoritmos diferentes. *b*) No alinhamento local, a prioridade é encontrar as regiões altamente similares, independentemente do tamanho desta região. Neste caso, porções da sequência que não foram alinhadas com alta similaridade foram excluídas do resultado final. *c*) No alinhamento global, as duas sequências são alinhadas por completo, independentemente do número de lacunas que tenham que ser inseridas.

milaridade entre elas pode ser computado, apresentamos três dos principais algoritmos desenvolvidos para este fim: algoritmos de programação dinâmica, análise de matriz de pontos (*dot matrix*) e método de palavra ou *k-tuple*.

A programação dinâmica é, atualmente, o método mais utilizado por programas para realizar o alinhamento de sequências. Em casos simples (par-a-par), é capaz de encontrar o melhor alinhamento para duas sequências através da aplicação da pontuação de similaridades. É, portanto, um método de execução relativamente rápida nos computadores modernos, requerendo um tempo e memória de processamento proporcional ao produto do tamanho das duas sequências envolvidas.

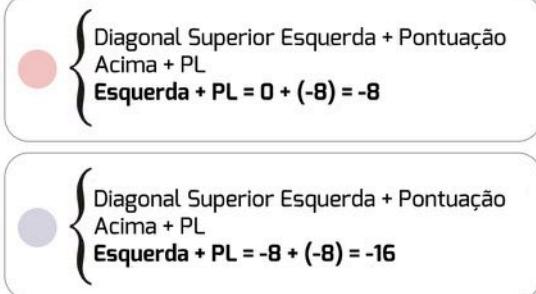
O método é baseado no princípio de otimização de Bellmann, e propõe a solução de problemas complexos através da resolução dos seus diversos subproblemas. Os subproblemas são resolvidos e seus resultados são armazenados pelo algoritmo. A vantagem funcional da resolução em partes é que, geralmente, problemas complexos combinam uma série de subproblemas. Como o algoritmo acumula os resultados dos diferentes subproblemas, acelera a resolução do problema complexo. Assim, a designação "programação" nada tem a ver com programação de computadores, mas com a organização dos resultados já solucionados para resolução de um problema maior.

Conforme discutimos anteriormente, em determinados casos, duas sequências podem apresentar diferentes alinhamentos. Se não há *indels* e as sequências são similares, o alinhamento é rápido e não deixa dúvidas. No entanto, quando existe certa diversidade entre as sequências envolvidas e uma quantidade suficiente de *indels*, a solução para o alinhamento é menos óbvia visualmente. Nestes casos, os algoritmos de programação dinâmica buscarão solucionar os subproblemas envolvidos e fornecerão o melhor resultado.

Para cálculo do melhor alinhamento entre duas sequências, o algoritmo de programação dinâmica necessita da especificação de um esquema de pontuação, seja ele referente a nucleotídeos ou aminoácidos. Da mesma forma, é necessário fornecer um valor de penalidade para a abertura e extensão das lacunas. A partir destas informações, o algoritmo calculará uma relação entre todos os caracteres das sequências e fornecerá o melhor alinhamento como resultado final.

Como exemplo, consideraremos a Figura 8-3. São dadas duas sequências, sequência 1 e sequência 2, um esquema de pontuação e, para facilitar o entendimento do cálculo, um valor único de penalidade por lacuna de -8. O algoritmo toma as sequências e transforma a relação entre elas em uma tabela, onde as linhas são definidas pelos caracteres da sequência 01, e as colunas pelos caracteres da sequência 02. A fim de permitir lacunas no início do alinhamento, o algoritmo impõe a inserção de uma coluna e de uma linha iniciais contendo o símbolo de *indel*. A partir deste ponto, para cada um dos elementos da matriz, o algoritmo calculará a melhor pontuação dos subcaminhos associados ao alinhamento: uma substituição, uma inserção na sequência 01 ou uma inserção na sequência 2. Assim, o melhor subcaminho será calculado segundo uma função de pontuação, conforme abaixo:

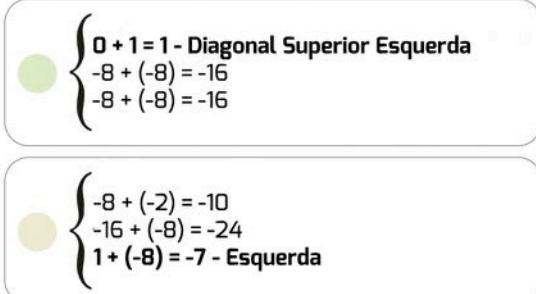
$$F(i,j) = max \begin{cases} \text{valor da célula na diagonal superior esquerda} + \text{pontuação da similaridade} \\ \text{valor da célula acima} + \text{valor da penalidade por lacuna} \\ \text{valor da célula à esquerda} + \text{valor da penalidade por lacuna} \end{cases}$$

A partir do elemento (1,1) da matriz e ao longo da primeira linha, apenas a terceira condição é satisfeita (valor da célula à esquerda + valor da penalidade por lacuna). Na primeira coluna, apenas a segunda condição é satisfeita. Para outros elementos, as três condições devem ser calculadas e aquela que resultar no maior valor é escolhida para formar a matriz. Além disso, os procedimentos dos algoritmos de programação dinâmica podem ser representados por pequenas setas para indicar qual subcaminho obteve o melhor valor (Figura 8-3).

Outro método importante na área de alinhamento de sequências é a análise de matriz de pontos ou matriz *dot*. É um método simples e bastante eficiente em análises de

a.

	–	A	C	C	G	T
–	0	-8	-16			
A						
G						
T						

Diagonal Superior Esquerda + Pontuação
Acima + PL
Esquerda + PL = 0 + (-8) = -8

Diagonal Superior Esquerda + Pontuação
Acima + PL
Esquerda + PL = -8 + (-8) = -16

b.

	–	A	C	C	G	T
–	0	-8	-16	-24	-32	-40
A	-8	1	-7			
G	-16					
T	-24					

0 + 1 = 1 - Diagonal Superior Esquerda
-8 + (-8) = -16
-8 + (-8) = -16

-8 + (-2) = -10
-16 + (-8) = -24
1 + (-8) = -7 - Esquerda

c.

	–	A	C	C	G	T
–	0	-8	-16	-24	-32	-40
A	-8	1	-7	-15	-23	-31
G	-16	-7	-1	-9	-14	-22
T	-24	-15	-9	-3	-11	-13

d.

Alinhamento Resultante:

Sequência 01 ACCGT

Sequência 02 A--GT

Caracter da Sequência 1, lacuna na sequência 2

Caracter da Sequência 2, lacuna na sequência 1

Caracteres das duas sequências

Figura 8-3: Alinhamento de duas sequências de nucleotídeos através do método de programação dinâmica. *a*) As sequências a serem alinhadas são dispostas em uma tabela onde o número de colunas corresponde ao número de caracteres da sequência 1 mais um (devido à adição de uma coluna para uma lacuna) e o número de linhas corresponde ao número de caracteres da sequência 2 mais um. O caractere atribuído à primeira linha e à primeira coluna é, por definição, o símbolo “-“, atribuído a uma lacuna. Através da matriz de penalidades calculam-se os valores para as três possibilidades $F(i,j)$, buscando a equação que resulte no maior valor. O valor arbitrário de penalidade por lacuna (*PL*) é de -8. Em virtude de a primeira linha não possuir valores de comparação na diagonal superior esquerda e acima, considera-se apenas a terceira equação. *b*) O valor demarcado em verde é o primeiro a ser calculado após o preenchimento da primeira linha e primeira coluna, representando o menor valor encontrado no cálculo para $F(i,j)$. Além do cálculo, o algoritmo de programação dinâmica insere informações a respeito da direção da informação. Como o valor “1” foi o maior valor encontrado e representa o cálculo utilizando a informação situada na diagonal superior esquerda, demarcada em verde, insere-se uma seta nesta direção. *c*) O preenchimento completo da tabela e as respectivas setas ilustrando a direção da informação. Algumas casas estão demarcadas com duas setas, pois apresentaram dois valores máximos idênticos na resolução das equações. Ao final dos cálculos, iniciando pelo canto inferior direito, seguem-se as setas em busca dos maiores valores. *d*) Relacionando os dados da tabela com a simbologia apresentada, chega-se ao alinhamento final entre as sequências 1 e 2.

deleções/inserções e para detectar repetições diretas ou inversas, especialmente em sequências de nucleotídeos. Além disso, vem sendo utilizado para buscar regiões de pareamentos intra-cadeia capazes de formar estruturas 2^árias^ em moléculas de RNA. Este método permite a visualização gráfica das regiões de similaridade entre sequências através da construção de uma matriz de identidade. O número de linhas desta matriz é definido pelo número de caracteres de uma das sequências, e o número de colunas é definido pelo número de caracteres da outra sequência a ser comparada (Figura 9-3). É primariamente um método visual, e não fornece o alinhamento propriamente dito como resultado final, embora seja frequentemente utilizado quando se deseja visualizar as regiões de similaridade entre duas sequências.

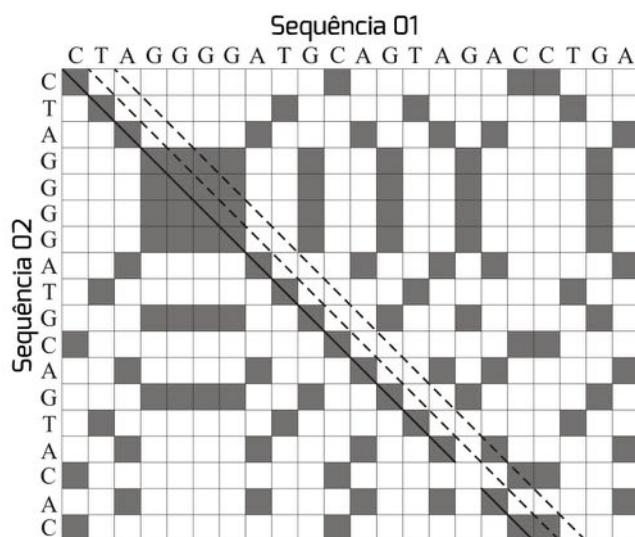



Figura 9-3: Análise de matriz de pontos de duas sequências de DNA. Os pontos assinalados em cinza representam a concordância de caracteres entre a sequência 1 e a sequência 2. A partir da diagonal direita inferior, são traçadas diferentes retas. Aquela que atingir o maior número de pontos assinalados deve ser escolhida como resultado para o alinhamento entre as duas sequências. A linha contínua representa a possibilidade mais adequada a esta análise e as linhas tracejadas representam possibilidades de insucesso.

Neste método, inicialmente, uma das sequências é disposta na vertical e a outra na horizontal (Figura 9-3). Regiões do gráfico que possuam o mesmo caractere tanto na sequência disposta na horizontal, quanto na sequência disposta na vertical, serão assinalados. Esta marcação representa os possíveis correspondências (*matches*) entre uma sequência e outra.

Qualquer região de similaridade entre as duas sequências será evidenciada por uma linha diagonal de assinalações. Pontos não dispostos na diagonal representam correspondências aleatórias que não estão relacionadas com a similaridade entre as sequências. A detecção de regiões de alta similaridade pode ser beneficiada, em alguns casos, através da comparação de dois ou mais caracteres ao mesmo tempo. Nestes casos, é necessário escolher um número de caracteres como janela.

Além disso, arbitrariamente, um número de correspondências deve ser escolhido. Por exemplo, para comparar duas sequências com 100.000 caracteres, podemos escolher uma janela de 15 caracteres e 10 correspondências requeridas. O algoritmo varrerá a matriz de 15 em 15 caracteres e, quando, entre estes quinze caracteres, existirem 10 formando correspondências entre as duas sequências, o algoritmo inserirá uma marcação de similaridade. Geralmente, esta variação do método é utilizada para a comparação de longas sequências de DNA.

Por último, outro algoritmo bastante comum no alinhamento par-a-par de dados biológicos é o *k-tuple*, ou método de palavras. Este método é geralmente mais rápido que o método de programação dinâmica, embora não garanta o melhor alinhamento como resultado. Este tipo de algoritmo é especialmente útil em casos onde se busca similaridade de uma única sequência contra um grande conjunto de dados. Para isso, o algoritmo dividirá uma sequência alvo em pequenas sequências, geralmente conjuntos de dois a seis caracteres, chamados de palavras. Da mesma forma, o conjunto total de sequências do banco de dados terá cada uma das sequências subdivida em pequenas pala-

vras. As palavras da sequência alvo serão comparadas às palavras oriundas do banco de dados. Após a busca de identidade, o algoritmo alinhará as duas sequências completas (sequência oriunda do banco de dados que teve uma palavra similar com umas das palavras da sequência alvo e a própria sequência alvo) a partir das palavras similares e estenderá a análise de similaridade para as regiões vizinhas, antes e depois da palavra similar. Através de uma matriz de penalidade, o algoritmo calculará o alinhamento que teve o maior valor de pontuação. É comum, para esta segunda etapa dos cálculos de similaridade, a utilização de algoritmos de programação dinâmica.

3.5. Alinhamento múltiplo global

Da mesma forma que no caso dos alinhamentos simples, o método de programação dinâmica é usualmente utilizado para lidar com múltiplas sequências. Nestes casos, utiliza-se o conceito de soma ponderada dos pares (*weighted sum of pairs*, WSP). Através deste conceito, para qualquer alinhamento múltiplo de sequências, uma pontuação para cada par possível formado por estas sequências será calculada (Figura 8-3) e, ao final, os valores de similaridade para cada um dos pares serão somados. Apesar de conceitualmente simples, este método exige grande capacidade computacional e, dependendo da quantidade de sequências envolvidas, pode requerer longo tempo para processamento.

Métodos alternativos tiveram que ser criados para acelerar os cálculos para alinhamento de sequências, incluindo-se: alinhamento progressivo, pontuação baseada em consistência (*consistency-based scoring*), métodos iterativos de refinamento, algoritmos genéticos e modelos ocultos de Markov. Cabe ressaltar que todos estes métodos realizam buscas aproximadas pelo resultado ótimo e, portanto, se tratam de métodos heurísticos.

Alinhamento progressivo

Leva em consideração a relação evolutiva entre as sequências. Os algoritmos utilizam as relações filogenéticas para gerar o resultado de alinhamento. Inicialmente, são realizados alinhamentos par-a-par de todos os possíveis pares. Nesta comparação, verifica-se apenas o número de caracteres diferentes entre as duas sequências (verificar o conceito de distância evolutiva observada no capítulo 6). Estas distâncias serão utilizadas para a construção de uma filogenia (geralmente através do método de *neighbor-joining*). A partir desta filogenia o alinhamento será construído progressivamente, dependendo da relação entre as sequências sendo, por isso, chamado de alinhamento progressivo.

Tomemos como exemplo um ramo de uma dada filogenia que inclui duas sequências. O algoritmo construirá um alinhamento através de programação dinâmica para estas duas sequências. A partir deste primeiro alinhamento, estas duas sequências serão agora tratadas como uma, e serão alinhadas à próxima sequência filogeneticamente relacionada. Devemos notar que todo o restante das sequências será alinhado baseando-se neste primeiro par. É um método rápido e amplamente utilizado para alinhar um grande número de sequências. Atualmente, os programas mais populares de alinhamento progressivo são o CLUSTALW e CLUSTALX.

Pontuação baseada em consistência

Baseado no algoritmo de alinhamento progressivo, não leva em consideração apenas o primeiro par de sequências alinhadas. Durante a realização do cálculo, realiza outros alinhamentos par-a-par para aperfeiçoar as comparações entre as sequências. O principal programa a utilizar este algoritmo é o T-COFFEE.

Métodos iterativos de refinamento

Funcionam como os algoritmos de ali-

nhamento progressivo, mas os grupos de sequências são realinhados constantemente ao longo das análises, garantindo que o alinhamento inicial não defina o resultado final. O principal programa a utilizar este algoritmo como base para os cálculos de alinhamento é o MUSCLE.

Algoritmos genéticos

Estes algoritmos buscam simular o processo evolutivo no conjunto de sequências a serem alinhadas, aplicando conceito de seleção e recombinação. É ainda um método lento e, devido à aleatoriedade do processo, não garante o mesmo resultado para diferentes alinhamentos do mesmo conjunto de dados. O programa SAGA é um dos poucos a implementar algoritmos genéticos.

Modelos ocultos de Markov

Modelo baseado em probabilidades estatísticas, destacando os eventos de substituição e inserção ou deleção de caracteres.

3.6. Alinhamento múltiplo local

Na busca por regiões localizadas de similaridade entre diferentes sequências, são aplicados principalmente os seguintes algoritmos: análise de perfis, análise de blocos e análise de motivos.

Análise de perfis

A partir de um alinhamento primário de todas as sequências envolvidas na análise e utilizando uma matriz de custo padrão, o algoritmo seleciona as regiões altamente conservadas e produz uma nova matriz de pontuação (matriz de custo), chamada de perfil. A construção deste perfil pode ser realizada através de dois métodos diferentes (método das médias e método evolutivo) e inclui pontuações para *matches*, *mismatches* e lacunas. Assim que produzido, este perfil pode ser utilizado para alinhar sequências entre si utilizando as pontuações calculadas para avaliar a probabilidade em cada posição ou para buscar sequências com o mesmo padrão em um banco de dados.

A desvantagem do método de perfis está na especificidade da nova matriz de custo obtida. Se o alinhamento inicial contiver poucas sequências, pode não representar adequadamente a variabilidade de caracteres em uma determinada posição e prejudicar o algoritmo na busca por similaridade com outras sequências. Este método é principalmente utilizado para alinhamentos de aminoácidos.

Análise de blocos

Assim como a análise de perfis este método requer, inicialmente, a seleção da região de maior similaridade de um alinhamento múltiplo. Estas regiões podem ser chamadas de blocos e diferem dos perfis por não acomodarem *indels*, que serão automaticamente eliminados das análises. Este método é também capaz de realizar a busca de pequenas regiões de similaridade entre sequências, de maneira semelhante ao método de palavras.

Análise de motivos

Este método é especialmente utilizado na busca por motivos proteicos em sequências de aminoácidos. O método foi desenvolvido através do alinhamento de milhares de sequências de aminoácidos extraídas de grandes bancos de dados de proteínas. A partir deste alinhamento, analisou-se cada uma das colunas para buscar um padrão de substituição entre os aminoácidos. Estes padrões de mudança refletem uma maior probabilidade de substituição. Para proceder ao alinhamento, os algoritmos que aplicam a análise de motivos iniciam o processo por uma análise de blocos. As regiões de alta similaridade são então analisadas para buscar os padrões de substituição descritos inicialmente. O conjunto de padrões resultante da análise das colunas é chamado de motivo. A probabilidade de existência de cada motivo em uma sequência de proteína é estimada através do banco de dados do SwissProt.

3.7. BLAST

O BLAST, ou Ferramenta de Busca por Alinhamento Local Básico (*Basic Local Alignment Search Tool*) é um algoritmo capaz de realizar buscas baseadas em alinhamento que, apesar de não serem exatas, são confiáveis e muito rápidas, sendo estas suas vantagens em relação a outros métodos. Ele é um dos programas mais usados em Bioinformática devido à velocidade em que consegue responder a um problema fundamental em biologia celular e molecular: comparar uma sequência desconhecida com aquelas depositadas em bancos de dados.

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências ao buscar primeiro por palavras comuns (ou k-*tuples*) na sequência de busca e em cada sequência do banco de dados. Em vez de buscar todas as palavras de mesmo tamanho, o BLAST limita a busca àquelas palavras que são mais significantes. O tamanho de palavra é fixado em 3 caracteres para sequências de aminoácidos e em 11 para sequências de nucleotídeos (3 se as sequências forem traduzidas nos 6 quadros de leitura possíveis). Esses são os tamanhos mínimos para obter uma pontuação por palavras que seja alta o suficiente para ser significativa sem perder fragmentos menores, mas importantes, de sequência.

Funcionamento do algoritmo BLAST

Para funcionar, o BLAST necessita de uma sequência de busca (*query*) e de sequências alvo. Comumente, as sequências alvos são o conjunto de sequências depositadas em um banco de dados, local ou na *web*. Um dos conceitos principais empregados pelo BLAST é de que alinhamentos estatisticamente significantes contêm pares de segmentos de alta pontuação (HSP, *high-scoring segment pairs*), e são esses HSPs que o algoritmo busca entre a sequência sendo analisada e aquelas depositadas no banco de dados.

As principais etapas do funcionamento do algoritmo BLAST, para uma sequência proteica genérica incluem:

i. Remoção de repetições ou regiões de baixa complexidade na sequência de busca.

Uma região de baixa complexidade é definida como uma região composta por poucos tipos de elementos. Essas regiões normalmente apresentam pontuações altas que podem confundir o programa em sua busca por sequências com similaridade significativa. Por esse motivo, tais regiões são identificadas antes da próxima etapa e ignoradas.

ii. Estabelecer uma lista de palavras com k-letras.

Sendo este um caso envolvendo sequências proteicas, k = 3, ou seja, cada palavra tem tamanho 3. Como mostrado na Figura 10-3, são listadas palavras com comprimento de 3 caracteres, sequencialmente, até que a última letra da sequência de busca seja incluída.

Figura 10-3: Exemplo de lista de palavras geradas pelo BLAST.

iii. Listar as possíveis palavras correspondentes.

Diferente de outros algoritmos (como o FASTA), o BLAST considera apenas as palavras de maior pontuação. As pontuações são estabelecidas por comparação das palavras listadas na etapa ii com todas as outras palavras de 3 letras. Uma matriz de substituição (BLOSUM62) é usada para pontuar as comparações entre pares de resíduos. Existem 20^3 possíveis pontuações de correspondência considerando uma palavra de 3 letras. Como exemplo, a comparação das palavras PQG e PEG tem pontuação de 15, enquanto a comparação de PQG com PQA pontua como 12. A seguir, um limiar T para pontuação de palavras vizinhas é usado para reduzir o número de possíveis palavras correspondentes. As palavras cujas pontuações forem maiores que o limiar T serão mantidas na lista de possíveis correspondências, enquanto aquelas cujas pontuações

forem menores serão descartadas. Considerando o exemplo anterior, se T = 13, PEG será mantida, enquanto PQA será abandonada.

iv. Organizar as palavras de alta pontuação.

As palavras remanescentes, com alta pontuação, são organizadas em uma árvore de busca. Isso permite que o programa compare as palavras com as sequências do banco de dados de maneira rápida.

v. Repetir os passos *iii* e *iv* para cada palavra de k-letras originadas da sequência de busca.

vi. Varrer as sequências do banco de dados em busca de correspondências com as palavras remanescentes.

O BLAST realiza uma varredura das sequências depositadas no banco de dados, buscando pelas palavras de alta pontuação (como PEG, no exemplo anterior). Se uma correspondência exata for encontrada, ela será empregada para nuclear um possível alinhamento sem lacunas (*gaps*) entre a sequência de busca e a depositada no banco de dados.

vii. Estender as correspondências exatas entre pares de segmentos de alta pontuação.

A versão original do BLAST estende o alinhamento para a esquerda e para a direita de onde ocorre uma correspondência exata. A extensão é parada apenas quando a pontuação acumulada pelo HSP começa a diminuir (um exemplo pode ser visto na Figura 11-3).

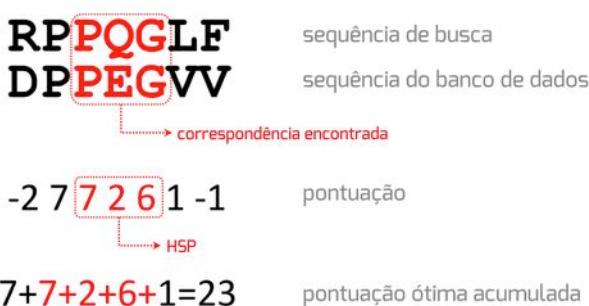



Figura 11-3: Exemplo do esquema de pontuação empregado pelo BLAST.

Para acelerar o processo, a versão atual do BLAST (BLAST2 ou *Gapped* BLAST) emprega um limiar mais baixo para a vizinhança das palavras, mantendo a sensibilidade na detecção de similaridade de sequências. Assim, a lista de possíveis correspondências obtidas na etapa *iii* é maior. Como observado na Figura 12-3, as regiões de correspondência exata com distância menor que A na mesma diagonal serão unidas como uma nova região, mais extensa. Posteriormente, essas regiões são estendidas da mesma maneira como ocorre no BLAST original, com os HSPs sendo pontuados com base em uma matriz de substituição.

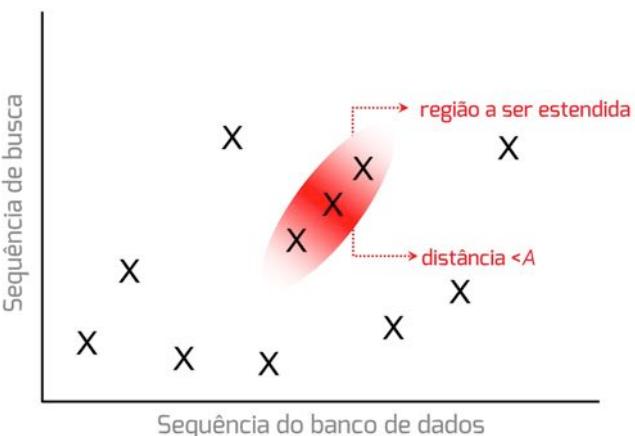



Figura 12-3: Esquema da extensão de zonas de correspondência entre sequências identificadas pelo BLAST.

viii. Listar todos os HSPs do banco de dados cuja pontuação seja alta o suficiente.

Nessa etapa são listados todos os pares de segmentos cuja pontuação seja maior que um determinado ponto de corte S. A distribuição de pontuações obtidas por alinhamento de sequências aleatórias é a base para determinação desse ponto de corte.

ix. Avaliar a significância da pontuação dos HSPs.

A avaliação estatística de cada par de segmentos de alta pontuação explora a Distribuição de Valores Extremos de Gumbel. O valor de confiança estatística e apresentado pelo BLAST, chamado de valor de expectativa, reflete o número de vezes que uma sequência não relacionada presente no banco de dados pode obter, ao acaso, um valor maior que S (ponto de corte). Ou seja, o e reflete o número de falsos positivos entre os resultados de similaridade encontrados. Para p < 0,1, o valor e se aproxima da distribuição de Poisson (ver item 4.8).

x. Transformar duas ou mais regiões de HSP em um alinhamento maior.

Em alguns casos, duas ou mais regiões de HSP podem ser combinadas em um trecho maior de alinhamento (uma evidência adicional da relação entre a

sequência de busca e a encontrada no banco de dados). Existem dois métodos para comparar a significância das novas regiões ligadas. Se, por exemplo, forem encontradas duas regiões de HSP combinadas com pares de pontuação (67 e 41) e (53 e 45), cada método se comportará de maneira diferente. O método de Poisson conferirá maior significância ao conjunto com valor mínimo maior (45 em vez de 41). O método de soma dos pontos, ao contrário, dará preferência ao primeiro conjunto, pois 108 (67+41) é maior que 98 (53+45). O BLAST original usa o primeiro método, enquanto o BLAST2 emprega o segundo.

xi. Exibir os alinhamentos locais entre a sequência de busca e cada uma das correspondências no banco de dados.

O BLAST original produz apenas alinhamentos sem lacunas (*gaps*), incluindo cada um dos HSPs encontrados inicialmente, mesmo que mais de uma região de correspondência seja encontrada numa mesma sequência do banco de dados. O BLAST2 produz um único alinhamento com lacunas, podendo incluir todas as regiões de HSP encontradas. É importante destacar que o cálculo da pontuação e do valor e leva em conta as penalidades por abertura de lacunas no alinhamento.

xii. Registrar as correspondências encontradas.

Quando o valor *e* dos alinhamentos encontrados entre a sequência de busca e as do banco de dados satisfazem o ponto de corte estabelecido pelo usuário, a correspondência é registrada. Os resultados da busca são apresentados de forma gráfica, seguidos por uma lista de correspondências organizada pela pontuação e pelo valor *e*, e finalizam com os alinhamentos. A Figura 13-3 traz um exemplo de resultado obtido pelo BLAST.

Diferentes tipos de BLAST

O BLAST constitui uma família de programas, que podem ser usados para diferentes fins, dependendo das necessidades do usuário. Esses programas variam quanto ao tipo de sequência de busca, o banco de dados a ser empregado, e o tipo de comparação a ser realizada. As diferentes aplicações disponíveis pelo BLAST incluem:

i. *blastn*: BLAST nucleotídeo-nucleotídeo. Usando uma sequência de DNA como entrada, dá como resultado as sequências de DNA mais similares presentes no banco de dados especificado pelo usuário.

ii. *blastp*: BLAST proteína-proteína. Usando uma sequência proteica como entrada, dá como resultado as sequências proteicas mais similares presentes no banco de dados especificado pelo usuário.

iii. *blastpgp*: BLAST iterativo com especificidade de posição (PSI-BLAST). Usado para encontrar proteínas distantemente relacionadas. Nesse caso, uma lista de proteínas proximamente relacionadas é criada. Essa lista serve de base para a criação de uma sequência média, que resume as características importantes do conjunto de sequências. A sequência média é usada para buscar sequências similares no banco de dados e um grupo maior de proteínas é encontrado. O grupo maior é usado na construção de uma nova sequência média e o processo é repetido. Ao incluir proteínas relacionadas na busca, o PSI-BLAST é muito mais sensível na percepção de relações evolutivas distantes que o BLAST proteína-proteína tradicional.

iv. *blastx*: tradução de nucleotídeos em 6 quadros-proteína. Compara os produtos de tradução conceitual nos 6 quadros de leitura de uma sequência de nucleotídeos contra o banco de dados de sequências proteicas.

v. *tblastx*: tradução de nucleotídeos em 6 quadros-tradução de nucleotídeos em 6 quadros. O mais lento dos programas BLAST, tem por objetivo encontrar relações distantes entre sequências de nucleotídeos. Ele traduz a sequência de nucleotídeo nos 6 possíveis quadros de leitura e compara os resultados contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco de dados.

vi. *tblastn*: proteína-tradução de nucleotídeos em 6 quadros. Compara uma sequência de proteína contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq.
Specific hits: Urease_gamma | Urease_beta
Superfamilies: Urease_gamma superfamily | Urease_beta superfamily
Multi-domains: PRK13986

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores: <40 | 40-50 | 50-80 | 80-200 | >=200

Query 1 40 80 120 160 200

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Urease subunit alpha; AltName: Full=Urea amidohydrolase subunit alpha >gb|AAA65722.1| urease [Helicobacter heilmannii]	475	475	100%	3e-168	100%	P42822.1
urease subunit beta [Helicobacter suis] >gb|EFX42255.1| Urease subunit alpha [Helicobacter suis HS5] >gb|EFX43059.1| Urease subunit alpha [Helic	441	441	100%	6e-155	92%	WP_006564485.1
UreA [Helicobacter bizzozeronii]	289	289	68%	4e-96	88%	ACR27088.1

Download GenPept Graphics ▼ Next ▲ Previous ▲ Descriptions

RecName: Full=Urease subunit alpha; AltName: Full=Urea amidohydrolase subunit alpha
Sequence ID: sp|P42822.1|URE23_HELHE Length: 234 Number of Matches: 1
▶ See 1 more title(s)

Related Information

Range 1: 1 to 234 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
475 bits(1222)	3e-168	Compositional matrix adjust.	234/234(100%)	234/234(100%)	0/234(0%)

```
Query  1    MKLTPKELDKLMLHYAGELAKQRKAKGIKLNYTEAVALISAHVMEEARAGKKSVADLMQE  60
            MKLTPKELDKLMLHYAGELAKQRKAKGIKLNYTEAVALISAHVMEEARAGKKSVADLMQE
Sbjct  1    MKLTPKELDKLMLHYAGELAKQRKAKGIKLNYTEAVALISAHVMEEARAGKKSVADLMQE  60

Query  61   GRTLLKADDVMPGVAHMIHEVGIEAGFPDGTKLVTIHTPVEAGSDKLAPGEVILKNEDIT  120
            GRTLLKADDVMPGVAHMIHEVGIEAGFPDGTKLVTIHTPVEAGSDKLAPGEVILKNEDIT
Sbjct  61   GRTLLKADDVMPGVAHMIHEVGIEAGFPDGTKLVTIHTPVEAGSDKLAPGEVILKNEDIT  120

Query  121  LNAGKHAVQLKVKNKGDRPVQVGSHFHFFEVNKLLDFDREKAYGKRLDIASGTAVRFEPG  180
            LNAGKHAVQLKVKNKGDRPVQVGSHFHFFEVNKLLDFDREKAYGKRLDIASGTAVRFEPG
Sbjct  121  LNAGKHAVQLKVKNKGDRPVQVGSHFHFFEVNKLLDFDREKAYGKRLDIASGTAVRFEPG  180

Query  181  EEKTVELIDIGGNKRIYGFNALVDRQADHDGKKLALKRAKEKHFGTINCGCDNK  234
            EEKTVELIDIGGNKRIYGFNALVDRQADHDGKKLALKRAKEKHFGTINCGCDNK
Sbjct  181  EEKTVELIDIGGNKRIYGFNALVDRQADHDGKKLALKRAKEKHFGTINCGCDNK  234
```

Figura 13-3: Exemplo de um resultado de busca realizada pelo BLAST. Diferentes informações são apresentadas: *1*) representação gráfica de domínios conservados identificados na sequência; *2*) representação gráfica de *matches*, indicando qualidade do alinhamento e cobertura das sequências identificadas; *3*) informações estatísticas dos resultados encontrados, incluindo identidade e valor *e*; *4*) alinhamento de cada sequência encontrada com a sequência de busca (*query*).

de dados.

vii. megablast: para empregar um grande número de sequências de busca. Quando se compara um grande número de sequências de busca (especialmente no BLAST por linha de comando), o megablast é muito mais rápido que o BLAST executado por várias vezes seguidas. Ele agrupa muitas sequências de busca, formando uma grande sequência, antes de realizar a busca no banco de dados. Os resultados são pós-analisados em busca de alinhamentos individuais.

3.8. Significância estatística

Em determinados casos, especialmente para buscar evidência de homologia entre sequências, o alinhamento é analisado sob o ponto de vista estatístico. Nessa óptica, podemos calcular quão bom pode ser um ali-

nhamento simplesmente levando em consideração as razões de chance de alinhamento entre nucleotídeos quaisquer. Para isso, sequências de nucleotídeos ou aminoácidos são geradas aleatoriamente, alinhadas em conjunto e avaliadas, segundo um determinado esquema de pontuação. Para alinhamentos globais, pouco se sabe a respeito destas distribuições randômicas. No entanto, felizmente, estas técnicas são bem entendidas para casos de alinhamentos locais e, atualmente, são amplamente utilizadas para a avaliação de similaridade, especialmente em bancos de dados que comportam grande quantidade de sequências.

Para analisar a probabilidade associada a determinado alinhamento é necessário, inicialmente, gerar um modelo aleatório das sequências em análise. Esses novos alinhamentos serão pontuados seguindo um determinado esquema de pontuação. Neste contexto, será calculada a probabilidade de se obter aleatoriamente uma pontuação pelo menos igual à pontuação do alinhamento original. O valor associado aos múltiplos testes realizados é chamado de valor e (e-*value*). Para banco de dados, este valor corresponde ao número de distintos alinhamentos, com uma pontuação igual ou melhor, que são esperados ocorrer na busca por sequências similares simplesmente por razões de chance (aleatórios). Estes cálculos estatísticos levam em consideração a pontuação do alinhamento e o tamanho do banco de dados. Quanto menor o valor e, menor o número de chances de uma determinada sequência ser alinhada aleatoriamente com outras e, portanto, mais significante é o resultado. Por exemplo, um valor e de 1e-3 ($1x10^{-3}$ ou 0,001) significa que há a chance de 0,001 de que a sequência alvo seja alinhada com uma sequência aleatória do banco de dados. Por exemplo, em um banco de dados que contém 10.000 sequências, neste caso, esperaríamos encontrar até 10 outras sequências que alinharão significativamente com a sequência alvo. É importante ressaltar que o fato de encontrarmos um valor e próximo de zero na comparação entre duas sequências não necessariamente denota a homologia destas sequências, dado que sequências não relacionadas podem conter similaridades devido à evolução convergente.

3.9. Alinhamento de 2 estruturas

O alinhamento de estruturas é um problema matematicamente complexo que só pode ser resolvido por algoritmos heurísticos. A Figura 14-3 apresenta um exemplo de alinhamento estrutural simples. Diferentes algoritmos oferecem resultados diferentes para o alinhamento, e algumas vezes essas diferenças são grandes. Por esse motivo é importante testar diferentes programas de alinhamento estrutural. Cada um deles tem pontos fortes e fracos, que podem ser explorados a partir da leitura dos artigos que os propuseram originalmente.

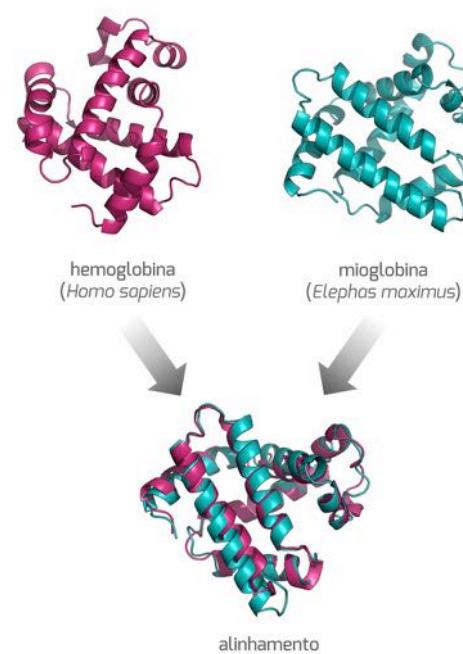



Figura 14-3: Exemplo de alinhamento de duas estruturas proteicas, oriundas de diferentes organismos: hemoglobina humana e mioglobina de elefante-asiático.

Existem três etapas essenciais para as diferentes estratégias de alinhamento estrutural: a representação, a otimização e a pontuação. A representação se refere às maneiras de representar as estruturas de uma forma que não seja dependente de coordenadas espaciais e que seja adequada ao ali-

nhamento. A otimização lida com a amostragem do espaço de possíveis soluções para o alinhamento entre as estruturas. A pontuação lida com a classificação dos resultados obtidos e com sua significância estatística. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de duas estruturas.

DALI: emprega matrizes de distâncias para representar as estruturas, transformando as estruturas 3D em conjuntos 2D de distâncias entre $C\alpha$. Se imaginarmos a sobreposição das matrizes, as regiões de sobreposição na diagonal representam similaridades na estrutura $2^{ária}$ (similaridades no esqueleto polipeptídico), e similaridades fora da diagonal representam similaridades na estrutura $3^{ária}$. As matrizes são então divididas em matrizes menores, de tamanho fixo, com base nas similaridades encontradas. Cada submatriz é unida a outras que sejam adjacentes para obter a matriz de sobreposição com maior abrangência. A significância estatística do alinhamento é calculada com base na distribuição encontrada em uma comparação de centenas de estruturas de baixa identidade. A pontuação é apresentada como número de desvios-padrão em relação a tal distribuição.

SSAP: cria vetores ligando resíduos a partir dos $C\beta$, representando a estrutura em duas dimensões, considerando posição e direção. Um algoritmo de programação dinâmica identifica similaridades entre as matrizes de vetores, gerando uma nova matriz que é posteriormente recalculada considerando as diferenças entre cada posição de similaridade encontrada na primeira etapa em relação às outras posições de similaridade, até que uma matriz ótima seja atingida. A pontuação do SSAP não é estatística, mas foi calibrada em relação ao banco de dados CATH. Assim, uma pontuação maior que 70 indica similaridade entre as estruturas comparadas.

VAST: cria vetores a partir de elementos de estrutura $2^{ária}$ cujo tipo, direção e conexão estão relacionados com a topologia da proteína. Esses elementos (fragmentos) de estrutura $2^{ária}$ são alinhados e comparados com alinhamentos gerados aleatoriamente. Alinhamentos com boa pontuação são agrupados e depois realinhados usando um procedimento de otimização por Monte Carlo. A significância estatística é dada pelo valor p (assim como ocorre no BLAST). O valor p é proporcional à probabilidade de se obter o alinhamento ao acaso.

SARF2: transforma as coordenadas em um conjunto de elementos de estrutura $2^{ária}$. Posteriormente, avalia pares desses elementos comparando o ângulo entre eles, a menor distância entre seus eixos e as distâncias mínimas e máximas entre cada elemento e a linha média. Um otimizador baseado em grafos é empregado para obter o maior número de conjuntos mutuamente compatíveis, e então o alinhamento final é calculado por adição de mais resíduos até que um valor mínimo de RMSD, definido pelo usuário, seja atingido. A pontuação final do alinhamento é calculada como função do RMSD e do número de $C\alpha$ pareados entre as estruturas. A significância estatística é obtida por comparação à distribuição de pontuações obtidas pelo alinhamento da proteína leghemoglobina a centenas de estruturas não redundantes.

CE: representa as proteínas como conjuntos de distâncias entre $C\alpha$ de oito resíduos consecutivos na estrutura. Primeiramente, são identificados todos os pares de octâmeros compatíveis entre as estruturas. Posteriormente, um algoritmo de extensão combinatória identifica e combina os pares mais similares entre as estruturas, adicionando mais pares a cada etapa do cálculo até a obtenção do melhor alinhamento. A significância estatística é dada por comparação às pontuações obtidas em um conjunto de alinhamentos entre estruturas com menos de 25% de identidade de sequência.

MAMMOTH: transforma as coordenadas da proteína em um conjunto de vetores unitários a partir dos $C\alpha$ de heptâmeros consecutivos. A similaridade entre heptâmeros é calculada pela sobreposição de seus vetores, a matriz de similaridade ótima é identificada e então o melhor alinhamento local entre estruturas é identificado dentro de um valor de RMSD pré-definido. A significância estatística é dada pelo valor p, baseado na comparação com a pontuação de alinhamentos obtidos aleatoriamente.

SALIGN: representa as proteínas por um conjunto de propriedades ou características calculadas a partir da sequência e da estrutura ou definidas arbitrariamente pelo usuário. Tais propriedades incluem tipo de resíduo, distância entre resíduos, acessibilidade da cadeia lateral, estrutura $2^{ária}$, conformação local da estrutura e característica a ser definida pelo usuário. O programa calcula uma matriz de dissimilaridade entre propriedades equivalentes, e a pontuação da dissimilaridade é calculada pela soma das matrizes de cada característica. A melhor sobreposição de matrizes é

obtida por um algoritmo baseado em programação dinâmica. A significância estatística não é calculada pelo SALIGN e o usuário obtém apenas os valores da pontuação de dissimilaridade. O programa fornece, entretanto, um valor adicional de qualidade, apresentado como porcentagem de Cα cuja distância é menor que 3,5 Å entre os pares de estruturas alinhadas.

3.10. Alinhamento de >2 estruturas

A maior parte dos métodos disponíveis para o alinhamento múltiplo de estruturas inicia-se estabelecendo todos os alinhamentos entre pares de estruturas e, então, emprega-os para estabelecer um alinhamento consenso entre todas as estruturas. A Figura 15-3 apresenta um exemplo de alinhamento estrutural múltiplo. Os métodos para obter o alinhamento consenso variam entre os programas de alinhamento. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de estruturas múltiplo.

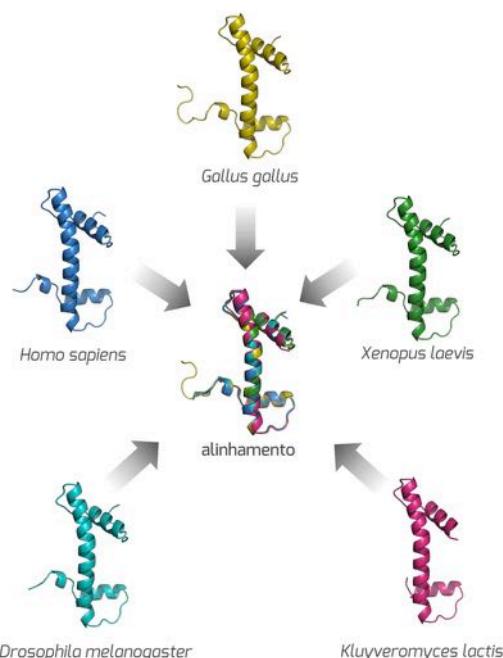



Figura 15-3: Exemplo de alinhamento de múltiplas estruturas proteicas, oriundas de diferentes organismos (histonas H3 de levedura, mosca-da-fruta, homem, frango, sapo-de-garras).

CE-MC: realiza o refinamento de um conjunto de alinhamentos de pares de estruturas empregando uma técnica de otimização de Monte Carlo. O algoritmo modifica o alinhamento múltiplo aleatoriamente, e as modificações são aceitas se houver melhoria na pontuação do alinhamento. O processo encerra quando o alinhamento múltiplo não puder mais ser melhorado por modificações aleatórias.

MAMMOTH-Mult: essa extensão do MAMMOTH gera inicialmente todos os alinhamentos de estruturas aos pares. Um procedimento de organização por médias é empregado para agrupar as estruturas com base em suas similaridades aos pares, gerando uma árvore. O alinhamento múltiplo é gerado por reorganização dessa árvore, onde ramos similares vão sendo agrupados aos pares, iterativamente.

SALIGN: pode realizar alinhamentos múltiplos de duas maneiras, baseado em uma árvore ou por alinhamento progressivo. O primeiro caso é muito similar ao MAMMOTH-Mult. No alinhamento progressivo, as estruturas são alinhadas na ordem em que são fornecidas para o programa. A vantagem desse método é o de seu custo computacional ser menor que o do método baseado em uma árvore.

3.11. Alinhamento flexível

O alinhamento de estruturas considerando sua flexibilidade está se tornando cada vez mais importante devido à melhor compreensão do enovelamento proteico. Cada vez mais, percebe-se que não existem enovelamentos estanques, mas sim um gradiente densamente populado por variantes conformacionais. Desta forma, torna-se mais difícil definir domínios proteicos, sendo mais adequado descrever as estruturas como conjuntos de estruturas supra-secundárias. Com base nessa proposta, a diferença entre proteínas relacionadas reside na orientação relativa desses subdomínios. A Figura 16-3 demonstra as diferenças que podem ser observadas ao alinhar um par de estruturas de maneira rígida ou flexível. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para este tipo de alinhamento de estruturas.

FATCAT: o algoritmo adiciona “torções” entre pares de fragmentos proteicos alinhados, que são tratados

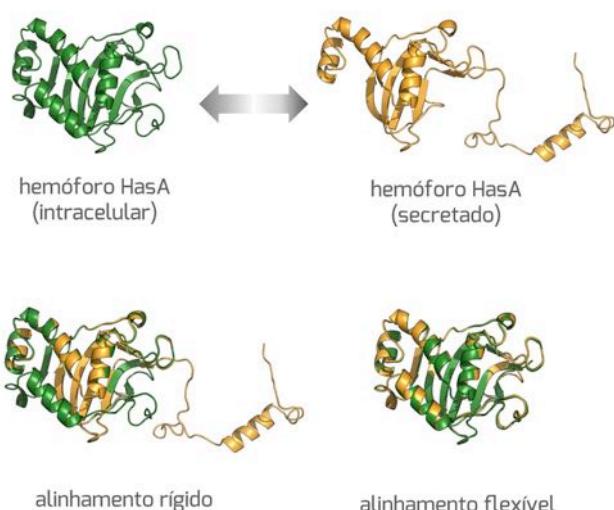



Figura 16-3: Comparação entre alinhamento estrutural rígido e flexível. A estrutura da proteína HasA (um captador bacteriano de grupamentos heme) foi obtida para suas formas intra- e extra-celular. Observe que o alinhamento rígido identifica similaridade parcial entre as estruturas, enquanto o alinhamento flexível detecta o rearranjo espacial de parte da proteína, evidenciando sua identidade.

como corpos rígidos. De maneira geral, o programa permite a inclusão dessas torções quando elas diminuem o valor final do RMSD, refletindo em um melhor alinhamento estrutural. O alinhamento final é obtido por programação dinâmica e se baseia na matriz de similaridade entre os fragmentos pareados, obtidos na primeira etapa do cálculo.

FLEXPROT: mantém uma das proteínas rígida, enquanto a outra pode sofrer alterações em busca de maior similaridade estrutural. As regiões potencialmente flexíveis da proteína são detectadas automaticamente e empregadas nas alterações conformacionais.

ALADYN: alinha pares de estruturas com base em sua dinâmica interna e similaridade entre seus movimentos de grande escala. O posicionamento ótimo entre as proteínas é encontrado ao maximizar as similaridades entre os padrões de flutuação estrutural, que são calculados pelo modelo de redes elásticas.

POSA: uma variante do FATCAT para o alinhamento múltiplo flexível de estruturas. Emprega uma metodologia combinada, introduzindo grafos de ordem parcial para visualizar e agrupar regiões similares entre as estruturas.

3.12. Conceitos-chave

Algoritmo: sequência lógica de instruções necessárias para executar uma tarefa.

Alinhamento: método de organização de sequências ou estruturas biológicas para evidenciar regiões similares e dissimilares. Estes métodos estão geralmente atrelados a inferências funcionais ou evolutivas.

Alinhamento Múltiplo: alinhamento que envolve mais de duas sequências ou estruturas

Alinhamento Simples: alinhamento que envolve apenas duas sequências ou estruturas.

BLAST: *Basic Local Alignment Search Tool* (Ferramenta de Busca por Alinhamento Local Básico), empregado para buscar sequências em bancos de dados com base em sua similaridade.

Homologia: é um termo essencialmente qualitativo que denota uma ancestralidade comum de determinada sequência.

HSP: pares de segmentos de alta pontuação (*high-scoring segment pairs*), zonas de similaridade entre sequências identificadas pelo BLAST.

Identidade: Porcentagem de caracteres similares entre duas sequências (excluindo-se as lacunas).

Indels: identifica inserções e deleções de caracteres ao longo do processo evolutivo.

Lacunas: regiões identificadas por hifens que representam a inserção/deleção de caracteres ao longo do processo evolutivo.

Matches: regiões que apresentam caracteres idênticos entre diferentes sequências.

Mismatches: regiões que apresentam caracteres não idênticos entre diferentes sequências.

Penalidades por lacuna (PL): conjunto de parâmetros necessários para atribuir a pontuação para uma lacuna em um sistema de alinhamento por pontuação.

RMSD: desvio médio quadrático.

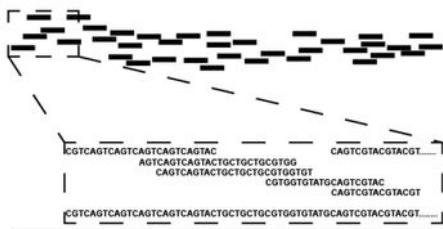
Tradução: tradução (*in silico*) de uma sequência de mRNA em sua possível sequência proteica correspondente

3.13. Leitura recomendada

BOGUSKI, Mark S. A molecular biologist visits Jurassic Park. ***Biotechniques***, 12, 668-669, 1992.

CARUGO, Oliviero. Recent progress in measuring structural similarity between proteins. ***Curr. Protein. Pept. Sci.***, 8, 219-241, 2007.

MADDEN, Tom. The BLAST sequence analysis tool. In: McENTYRE, Jo; OSTELL, Jim (Org.). ***The NCBI Handbook***. Bethesda: National Center for Biotechnology Information, 2002.

MARTI-RENOM, Marc A.; et al. Structure comparison and alignment. In: GU, Jenny; BOURNE, Philip E. (Org.). ***Structural Bioinformatics***. 2.ed. Hoboken: John Wiley & Sons, 2009.

MAYR, Gabriele; DOMINGUES, Francisco S.; LACKNER, Peter. Comparative analysis of protein structure alignments. ***BMC Struct. Biol***., 7, 50, 2007.

MOUNT, David W. ***Bioinformatics: Sequence and Genome Analysis***. 2.ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.

ROSSMANN, Michael G.; ARGOS, Patrick. The taxonomy of binding sites in proteins. ***Mol. Cell. Biochem.***, 21, 161-182, 1978.

Charley Christian Staats
Guilherme Loss de Morais
Rogério Margis

Representação da montagem de genomas.

4.1. Introdução

A análise *in silico* das sequências nucleotídicas de cromossomo(s) de um dado organismo, ou simplesmente genoma, constitui uma da mais importantes aplicações da bioinformática. Tem como objetivo desenvolver e utilizar ferramentas para identificar e caracterizar genes, elementos genéticos móveis e outros elementos presentes em um determinado genoma, assim como fazer intercorrelações entre diferentes genomas com o intuito de buscar aspectos evolutivos comuns.

O primeiro organismo a ter a sequência de nucleotídeos de seu genoma determinado foi a bactéria Gram negativa *Haemophilus influenzae*, em um projeto liderado por J. Craig Venter. Desde 1995, ano de publicação desta análise genômica, as sequências de milhares de genomas de outros organismos já foram determinadas e analisadas, não apenas de espécies, mas também de variedades de espécies, raças e linhagens, entre outros.

Com a grande disseminação de estratégias de sequenciamento cada vez menos onerosas, muito tem se investido na geração de algoritmos e programas para analisar as sequências genômicas geradas. Previamente às análises do genoma de *H. influenzae*, programas para montagem de genomas já existiam, tendo sido desenvolvidos para análise de volumes de sequências relativamente pequenos, como os dos fagos λ e CMV, com tamanhos de aproximadamente 48.000 pares de bases (pb) e 229.000 pb, respectivamente. Para genomas maiores, novos programas tiveram que ser desenvolvidos em virtude da maior complexidade e quantidade das sequências analisadas. Neste capítulo, serão abordados os conceitos básicos e as principais ferramentas para montagem e anotação de genomas, assim como alguns programas para a sua análise.

4.2. Montagem de genomas

Nos primeiros anos da era genômica, o sequenciamento de genomas era baseado na metodologia de Sanger, ou método dideóxi. Para obtenção da sequência dos genomas, os fragmentos de DNA gerados após fragmentação química, física ou enzimática eram subclonados em vetores plasmidiais. Esta estratégia, denominada sequenciamento *shotgun*, é baseada na fragmentação aleatória dos cromossomos em fragmentos de DNA com tamanho relativamente pequeno. Estes fragmentos, cujo tamanho geralmente variava de 2.000 a 5.000 pb, eram submetidos ao sequenciamento. As sequências obtidas a partir de cada clone (chamadas de *reads*), com tamanho médio de 600 a 800 pb, eram submetidos a um processamento para retirada de sequências de baixa qualidade e, então,

utilizadas na montagem de *contigs* e genomas (ver abaixo).

Com o advento das metodologias denominadas *next-generation sequencing* – NGS (pirossequenciamento, Illumina, SOLiD, dentre outros), também ocorre fragmentação aleatória do DNA genômico, mas geralmente não são necessários os passos de clonagem. Comparativamente, estes novos métodos permitem a obtenção de *reads* de maneira muito mais rápida. Entretanto, o tamanho dos *reads* é menor, variando de algumas dezenas a poucas centenas de pares de base, dependendo da metodologia. Assim como no sequenciamento por Sanger, os *reads* obtidos passam por um controle de qualidade e então podem ser utilizados na montagem de genomas.

Independente da metodologia de sequenciamento utilizada, como resultado se tem uma grande lista de sequências nucleotídicas - os *reads* - de tamanhos que podem variar de 50 a 800 pb. Para montagem das sequências genômicas a partir destes *reads*, diferentes estratégias são utilizadas, dependendo da metodologia empregada. Para o sequenciamento convencional (Sanger), cada um destes reads é alinhado entre si na procura de regiões de identidade ou de sobreposição, de maneira a construir fragmentos contíguos (*contigs*), os quais podem ser definidos como a união de duas ou mais sequências (*reads*) formadas por sobreposição de elementos comuns a pelo menos duas sequências (Figura 1-4).

Os primeiros algoritmos para montagem de genomas se baseavam no alinhamento dos *reads* e na concatenação de sequências obtidas dos *reads* com os maiores alinhamentos. O processo se dava de forma cíclica, concatenando as sequências com o maior alinhamento até que todos estes alinhamentos fossem utilizados. Esta montagem de genomas a partir de *reads* tem como base os seguintes passos:

i) cálculo de alinhamentos aos pares de todos os fragmentos;

ii) escolha de dois fragmentos com a maior sobreposição;

iii) fusão dos dois fragmentos;

iv) repetição dos passos anteriores até obtenção de uma única sequência.

Para as novas metodologias de sequenciamento, devido ao tamanho relativamente menor dos fragmentos, algoritmos diferentes foram desenvolvidos. Os

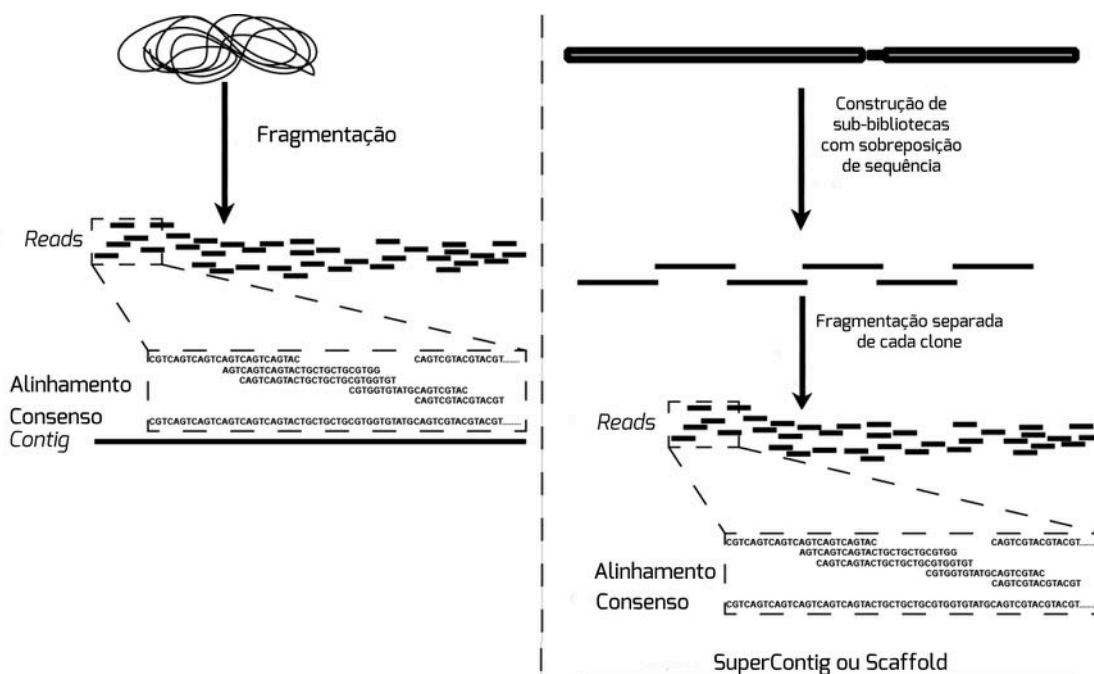



Figura 1-4: Montagem de genomas utilizando a estratégia de sequenciamento de genomas por *shotgun*. O painel à esquerda ilustra um esquema utilizado para genomas de menor tamanho e reduzido conteúdo de sequências repetitivas. O painel à direita ilustra uma estratégia mais complexa, usado para organismos com genoma maior.

programas de montagem atuais utilizam grafos de sobreposição ou grafos de Bruijn. Estes grafos identificam *reads* com possibilidade de compartilharem trechos de sobreposição entre si utilizando uma estratégia baseada no alinhamento em sementes.

Com esta abordagem, pequenos fragmentos de comprimento fixo obtido de cada *read*, os k-*mers*, são usados como um índice, e apenas pares de leituras que partilham uma semente são posteriormente avaliados. Os grafos de Bruiijn baseiam-se na decomposição de reads em k-*mers* (por exemplo dodecâmeros, ou seja fragmentos de 12 nucleotídeos), os quais são utilizados como nodos destes grafos. Uma ligação direta entre os nodos indica que estes k-*mers* ocorrem consecutivamente em um ou mais *reads*.

Uma série de programas foram desenvolvidos para a montagem de genomas, utilizando diferentes algoritmos (Tabela 1-4). No caso de sequenciamento de genomas procarióticos, ao final do processo é esperada a obtenção de uma sequência única, a qual representa toda a sequência nucleotídica do cromossomo. Sabe-se, todavia, que plasmídeos podem ser encontrados em diversos micro-organismos. Assim o número de *contigs* será dependente do número de plasmídeos e, em casos menos frequentes, do número de cromossomos presentes naquela bactéria.

Ao ser analisado o genoma de organismos eucariotos, nos quais se encontra uma grande variação no número de cromossomos, um número maior de *contigs* é esperado. Teoricamente, cada cromossomo deveria ser representado por um *contig*. Entretanto, nos passos iniciais de montagem de genomas são observados dezenas a centenas de *contigs*, dependendo da complexidade do organismo cujo genoma esta sendo sequenciado. Os genomas de eucariotos, em especial de eucariotos superiores, possuem pelo menos duas características que tornam o processo de montagem mais complexo:

i) uma quantidade considerável de sequências repetitivas que dificulta o processo de montagem devido a alinhamentos de alto escore com diversas sequências;

ii) o seu tamanho, podendo chegar a mais de 3 bilhões de pares de base (caso do genoma humano).

Tabela 1-4: Principais programas utilizados na montagem de genomas e transcriptomas.

Nome	Análise
ABySS	grandes genomas
ALLPATHS-LG	grandes genomas
Celera WGS *Assembler*	grandes genomas
CLC *Genomics Workbench*	genomas e trancriptomas
Geneious	genomas
Newbler	genomas e transcriptomas
Phrap	genomas e transcriptomas
SOAPdenovo	genomas e transcriptomas
Staden gap4 *package*	genomas pequenos e transcriptomas
Trans-ABySS	transcriptomas
Velvet	genomas pequenos e transcriptomas

Para sobrepujar estas dificuldades, passos intermediários se tornam necessários, como a construção de sub-bibliotecas genômicas. Cada uma destas sub-bibliotecas é sequenciada, de forma a gerar *contigs*. O conjunto de diferentes *contigs* oriundos de diferentes sub-bibliotecas será utilizado para a geração de *scaffolds* (Figura 1-4). Geralmente, são necessários passos adicionais de clonagens de regiões específicas do genoma e posterior sequenciamento destas para o "fechamento" do genoma.

Um dos maiores desafios, entretanto, para o sequenciamento de genomas reside na adequada montagem de regiões repetitivas. No genoma humano, por exemplo, existem pelo menos seis classes de sequências repetitivas:

i) minissatélites, microssatélites ou satélites;

ii) SINEs (elementos nucleares pequenos intercalados);

iii) LINEs (elementos nucleares longos intercalados);

iv) transposons;

v) retrotransposons;
vi) *clusters* de genes DNAr (genes responsáveis pela síntese dos RNA ribossômicos – RNAr).

Estas diferentes classes, cujos tamanhos podem variar de centenas de pares de base, caso de microssatélites e SINEs, a dezenas de milhares de pares de base, observado em *clusters* de genes DNAr, podem constituir mais de 50 % do tamanho de cada cromossomo humano.

O grande desafio na montagem de sequências genômicas com alto conteúdo de elementos repetitivos se refere a correta quantificação e localização destes elementos nos cromossomos. Desta forma, o desafio central da montagem de genomas reside na resolução destas sequências repetitivas, estando este desafio diretamente associado à metodologia de sequenciamento utilizada. Por exemplo, se forem obtidos *reads* de tamanho menor que uma unidade de repetição, todos estes *reads* serão utilizados para formar um *contig* que contém apenas a sequência de repetição. Entretanto, ao serem obtidos *reads* com tamanho maior que a unidade de repetição, os mesmos podem ser utilizados na resolução da localização destas sequências repetitivas em um determinado cromossomo.

Alguns programas permitem montar genomas complexos com repetições baseados em *reads* maiores (como os obtidos pela metodologia de Sanger ou pirosequenciamento). Para tal, estes programas realizam a montagem em duas ou mais fases distintas, nas quais as sequências repetitivas são processadas separadamente. Em um primeira fase do processo de montagem, *reads* contendo sobreposição de sequências não ambíguas são agrupados em *contigs*, cujas extremidades contém as regiões limítrofes das sequências de repetição. A segunda fase se caracteriza pela montagem de *contigs* não ambíguos em sequências maiores, usando dados de *reads mate-pair*.

Dados de sequenciamento *paired-end* oferecem a possibilidade da determinação exata de sequências que flanqueiam uma determinada sequência de repetição. Em experimentos tradicionais associados ao sequenciamento de Sanger, um protocolo *paired-end* inicia-se com longos fragmentos de DNA clonados em vetores para sua replicação em *Escherichia coli*. As extremidades destes fragmentos poderiam assim ser facilmente determinadas por sequenciamento. Protocolos *paired-end* para as estratégias de sequenciamento atuais não requerem passos de clonagem em *E. coli*. Entretanto, os mesmos se baseiam na circularização do fragmento de DNA do tamanho desejado, sendo as extremidades posteriormente reconhecidas devido à etiqueta (*tag*) utilizada para propiciar a circularização por meio da ligação. Com a determinação das sequências flanqueadoras de uma repetição, há maior chance de conseguir determinar a sua localização em um genoma.

A qualidade de montagem do genoma pode ser acompanhado por alguns índices. A cobertura reflete a quantidade de *reads* associados a um determinado fragmento de DNA. Por exemplo, uma cobertura de 10X indica que, para o genoma sendo avaliado, cada nucleotídeo foi encontrado em pelo menos 10 *reads*.

Outro valor importante refere-se ao N50. Trata-se de uma medida estatística muito utilizada para avaliar a qualidade da montagem, visto que revela o quanto de um genoma é coberto por *contigs* grandes. Um valor de N50 igual a n significa que 50% dos reads estão montados em um *contig* de tamanho n ou maior. Por exemplo, na montagem do genoma de cão doméstico, depositado no NCBI sob o número de acesso AAEX03, o sequenciamento dos 40 cromossomos, com uma sequência total de 2.410.976.875 bases gerou 27.106 *contigs* com um N50 de 267.678. Isto significa que mais de 50% dos *reads* estão associados a *contigs* de 267.678 bases ou maiores.

4.3. Montagem de transcriptomas

Em análises de novos genomas, um ponto importante se refere à identificação de transcritos. Além de fornecer indícios sobre quais genes estão sendo expressos em uma determinada situação fisiológica a qual as células ou tecidos estão sendo expostos, o sequenciamento de transcritos tem uma aplicação importante na procura de sequências codificantes em genomas. Esta estratégia tem uma aplicabilidade muito grande em organismos em que o conteúdo de íntrons por gene é grande, como em eucariotos mais complexos.

Ao contrário de genomas, em transcriptomas o material de partida geralmente é

cDNA, obtido a partir de transcrição reversa de RNA. A grande maioria dos trabalhos se dá em torno de RNAm mas, cada vez mais, RNAs não codificantes, com possível papel regulatório, estão sendo avaliados por esta metodologia (ver abaixo). O *pool* de cDNAs pode então ser subclonado e ser submetido ao sequenciamento pela metodologia de Sanger ou diretamente fragmentado e ser submetido ao sequenciamento NGS. Uma grande lista de *reads* é então obtida, os quais podem ser utilizados para realizar a montagem do transcriptoma *de novo* ou ser ancorados a sequência de um genoma para ajudar na identificação de sequências codificantes e de extremidades éxon/íntron.

No caso da montagem *de novo*, os *reads* são alinhados e aqueles que apresentam alinhamento positivo são fusionados, dando origem a *contigs*. Entretanto, diferentemente da análise de genomas, muitos *contigs* são gerados, cada um possivelmente representando um mRNA maduro.

Adicionalmente, alguns programas podem, além de realizar a montagem de transcriptomas ou alinhamento a genomas, fazer uma análise da representatividade de cada transcrito dentro do conjunto total de RNA analisado, por meio do cálculo da frequência relativa de cada transcrito identificado. Com estes cálculos é possível realizar análises de expressão diferencial de genes. Dentre os pacotes de programas utilizados, podem ser citados Cufflinks-Cuffdiff, DegSeq, DESeq, EdgeR, entre outros.

A análise desta expressão relativa de transcritos pode ser realizada com base em duas estratégias principais:

i) mapeamento a uma sequência genômica previamente conhecida;

ii) análise *de novo*, independente da sequência genômica e baseada na montagem dos transcritos diretamente a partir dos *reads*.

Na primeira estratégia, os *reads* são mapeados ao genoma, ou seja, as regiões de identidade nucleotídica são ancoradas à sequência genômica, sendo identificadas por metodologias de sequenciamento que levam em consideração o número de *reads* mapeados em relação à porção do genoma que contém um gene. Alguns dos programas para este tipo de mapeamento incluem Bowtie, Tophat e SOAP, dentre outros. Como resultado, uma determinada sequência do genoma é representada por um grande número de *reads*, no caso de genes mais expressos, ou um baixo número de *reads*, no caso de genes menos expressos.

Deve ser levado em consideração, entretanto, que quanto maior o tamanho do gene mais se espera encontrar *reads* associados a este gene. Desta forma, a maneira mais comum para se calcular a expressão relativa de um determinado gene é o RPKM (*reads per kilobase of transcript per million mapped reads* – *reads* por kilobase de transcrito por milhões de *reads* mapeados). Esta abordagem permite uma análise comparativa baseada em uma série de análises estatísticas para comparação de transcritos com diferentes RPKMs de diferentes amostras biológicas ou diferentes tempos de tratamento, por exemplo.

Quando são considerados organismos cujo genoma ainda não foi determinado, uma construção do transcriptoma a partir de dados de RNAseq é realizada (*de novo*). A partir das sequências dos transcritos gerados, é possível então fazer o cálculo do RPKM de cada transcrito identificado.

4.4. Identificação/anotação gênica

A anotação de genomas é o passo seguinte à montagem dos genomas. Trata-se de um conjunto de protocolos e fluxos de trabalho utilizados para delimitar, em uma determinada sequência genômica, possíveis genes e predizer a sua função com base na similaridade com sequências conservadas. Basicamente, existem dois grande grupos de genes avaliados nestas metodologias. O primeiro grupo se refere àqueles cujo produto é reconhecido pelos ribossomos e dará origem a uma proteína (ou seja, RNAm). Já o segundo engloba os genes cujo produto terá funções estruturais e funcionais dependentes da própria molécula de RNA, como RNAt e RNAr. Diferentes abordagens são utilizadas para identificar as sequências de cada um destes grupos de genes, como será visto abaixo.

Identificação de regiões codificantes

O mecanismo de delimitação da sequência gênica é drasticamente influenciado pelo Domínio ao qual pertence o organismo cuja sequência genômica foi determinada. Isto se deve ao fato de que existe uma grande diferença nas estruturas de genes procarióticos e eucarióticos.

Genes procarióticos codificantes de proteínas são colineares com seus produtos gênicos. Esta característica permite inferir que toda região delimitada por um códon de início e um códon de término, região esta denominada de ORF (*Open Reading Frame*), potencialmente constitui uma região codificante de uma proteína em um genoma procariótico.

Por sua vez, genes eucarióticos codificantes de proteínas são mais complexos, geralmente sendo caracterizados pela presença de sequências intervenientes ou íntrons. Até pouco tempo, acreditava-se que íntrons constituíam um produto da evolução que povoou as sequências gênicas com o chamado "DNA lixo", de modo que uma mutação que eventualmente viesse a acontecer tivesse maior possibilidade de ocorrer em regiões do gene que não têm capacidade codificante. Recentemente, contudo, determinou-se que os íntrons exercem um importante papel regulatório na expressão gênica.

Íntrons são elementos gênicos que, durante o processo de expressão gênica, são excisados durante o processamento do RNA, em um grande complexo de reações denominado *splicing*. Os íntrons podem variar em número e tamanho, dependendo da complexidade do organismo. Assim, em organismos mais simples, como leveduras e fungos filamentosos, o número de íntrons por gene é pequeno (geralmente de 1 a 4 por gene), assim como o seu tamanho (geralmente girando em torno de 50 pb).

Ao contrário, em organismos mais complexos como humanos e plantas, tanto o número de íntrons por gene quanto o seu tamanho aumentam significativamente, de forma que grande parte do gene é constituído por íntrons (mais de 90%, dependendo do organismo). Um comparativo entre as estruturas básicas de genes codificantes de proteínas procarióticos e eucarióticos, assim como os seus respectivos processos de expressão, é apresentado na Figura 2-4.

Associado ao grande número de íntrons, genes de organismos eucarióticos mais complexos geralmente são caracterizados pelo

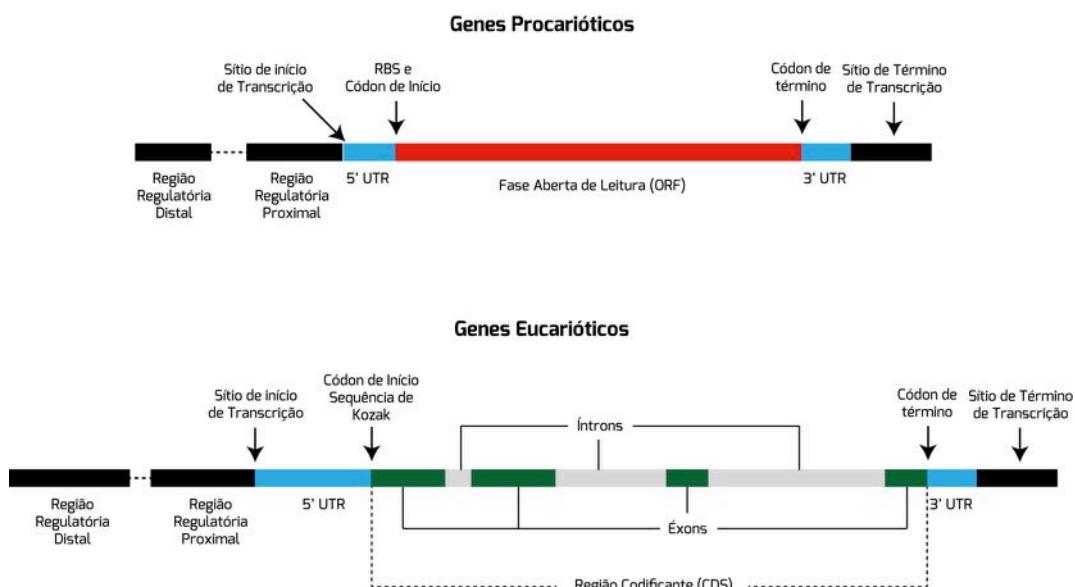



Figura 2-4: Esquema representando os elementos encontrados em genes procarióticos (quadro superior) e eucarióticos (quadro inferior). Os genes estão representados no sentido 5'-3' e podem ser notadas as principais diferenças entre estas classes de genes, como a presença de íntrons e regiões regulatórias mais complexas em eucariotos.

splicing alternativo. Este processo é caracterizado pela incorporação diferencial de íntrons e éxons no RNAm maduro, de forma a produzir diferentes proteínas a partir do mesmo gene.

Diferentes estratégias para procura de genes em genomas foram desenvolvidas considerando estas características diferenciais na estrutura de genes procarióticos e eucarióticos. A procura de ORFs em genomas procarióticos constitui uma estratégia simples e direta. Entretanto, é uma estratégia sujeita a uma diversidade de erros.

Nestas predições, não são considerados elementos canônicos clássicos presentes na estrutura de genes (isto é, sequências conservadas para ligação do fator sigma, região de ligação do ribossomo, sítio de início de tradução e sítio de término de tradução) e operons, os quais poderiam auxiliar na procura *ab initio* (ou seja, diretamente a partir de sequência, sem informações experimentais diretas sobre o produto gênico) de genes em genomas procarióticos. Assim, a procura de genes baseada apenas na identificação de ORFs geralmente leva a um número grande de resultados falsos positivos e falsos negativos (Figura 3-4).

Para sobrepujar estas limitações, mecanismos de delimitação das sequências gênicas em genomas procarióticos foram então desenvolvidos e se baseiam em algoritmos característicos para detectar, na sequência de DNA, dois tipos fundamentais de informações: sinais e conteúdo. Estes mecanismos foram então expandidos para procura de genes em organismos eucarióticos.

Os detectores de sinais procuram por caracteres funcionais específicos de genes, tanto associados à transcrição quanto à tradução. Sinais transcricionais incluem sequências canônicas conservadas que delimitam as regiões necessárias para que se inicie o processo de transcrição. Os sinais mais comumente descritos em procariotos são as regiões -35 e -10 e as sequências de associação com a RNA Polimerase. Já os sinais procurados em sequências eucarióticas geralmente constituem a região TATA box, assim como o sítio de clivagem e poliadenilação, que caracteriza o terminador.

Os sinais traducionais, por sua vez, se referem basicamente às regiões importantes para recrutamento de ribossomos, como o RBS (*ribosome binding site*, ou sitio de ligação a ribossomos) em procariotos. Como este mecanismo é diferente em organismos eucarióticos, uma região conservada, denominada sequência de Kozak, é utilizada como sinal traducional em eucariotos. Estas duas regiões se localizam imediatamente a montante (*upstream*) aos respectivos códons de início, e desempenham um papel importante nos mecanismos de delimitação de genes.

Adicionalmente, a detecção de sinais que delimitam os íntrons também são utilizados pois, como abordado anteriormente, os genes de eucariotos são amplamente povoados por íntrons. Desta forma, a correta predição da posição de íntrons é fundamental para correta anotação do gene, sendo que os principais sinais a serem avaliados são os nu-

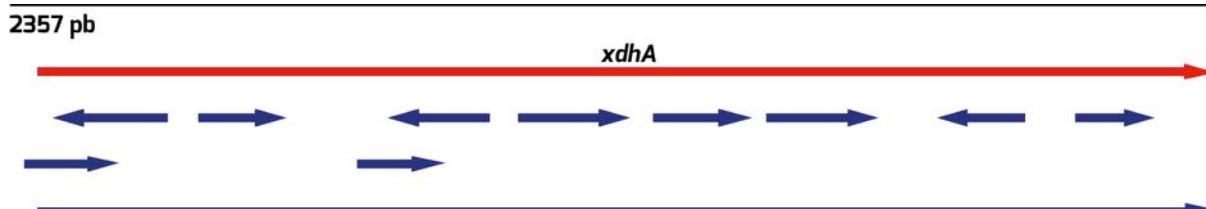



Figura 3-4: A simples procura de ORFs pode gerar resultados falso positivos na procura de genes em organismos procarióticos. Como exemplo, uma sequência de DNA de 2357 pb da bactéria *E. coli* HS (nucleotídeos 3027764 ao 3030120 – Código de Acesso junto ao NCBI NC_009800.1), o qual contém o gene xdhA, foi avaliada quanto à presença de ORFs com mais de 150 pb com o programa ORF Finder. A sequência anotada do gene encontra-se em vermelho, ao passo que as possíveis ORFs estão demarcadas em azul.

cleotídeos que compõem as extremidades conservadas 5' e 3' do íntron, mais comumente GT e AG (ver abaixo).

Já os detectores de conteúdo classificam a sequência de DNA em codificante e não-codificante. Como região não-codificante entendem-se íntrons, regiões intergênicas e regiões não traduzidas dos genes. Os detectores de conteúdo podem ainda ser subdivididos em detectores extrínsecos e detectores intrínsecos. Os detectores de conteúdo extrínsecos se baseiam no fato de que regiões codificantes são mais conservadas em relação às não-codificantes propiciando, desta forma, a identificação de éxons conservados com base em procuras por homologia.

O mecanismo básico desta busca é através do programa BLAST (ver capítulo 3). Contudo, uma limitação nesta metodologia se refere à avaliação adequada da presença de ortólogos diretos. Desta forma, a distância filogenética (isto é, evolutiva, ver capítulo 5) entre o organismo cujo genoma está sendo analisado e aqueles organismos cujas sequências estão depositadas nos bancos de dados pode influenciar diretamente no resultado.

Detectores de conteúdo intrínseco, por sua vez, tem como foco principal algumas características inatas do DNA, as quais permitem a predição do potencial de uma sequência codificar ou não uma proteína. Como exemplos de características avaliadas em detectores intrínsecos podem ser citados:

i) em muitos organismos há uma preferência das bases G ou C em relação às bases A ou T na terceira posição do códon;

ii) a utilização diferencial de códons sinônimos, ou seja, diferentes códons que codificam para o mesmo aminoácido;

iii) frequência de distintas sequências nucleotídicas hexaméricas;

iv) a periodicidade de ocorrência de bases, dentre outros.

Estes caracteres são utilizados, por exemplo, em modelos de Markov para a construção de modelos capazes de reconhecer sequências codificantes. Com base nos mecanismos discutidos acima, dois principais sistemas para procura de genes em genomas de eucariotos foram construídos, denominados empírico e *ab initio*.

Procura empírica de genes

A predição empírica ou baseada em evidência leva em consideração buscas por similaridade com outros bancos de dados (genômicos, transcritômicos ou proteômicos) para identificar e delimitar as sequências gênicas. Métodos de identificação de genes baseados em similaridade são considerados de alta confiabilidade para localizar e construir modelos gênicos, desde que existam relatos prévios de estruturas gênicas do próprio organismo (como, por exemplo, sequências de RNAm) ou baseado em análises de conservação provenientes de alinhamentos de genomas de espécies filogeneticamente relacionadas.

Especialmente para o caso de organismos eucarióticos, alinhamentos de sequências oriundas de bancos de dados de proteínas ou de transcritos contra o genoma em anotação permitem aferir que, geralmente, os *gaps* constituem os íntrons. Esta premissa é frequentemente acompanhada pela observação de que as sequências limítrofes dos íntrons identificados constituem os dinucleotídeos consenso GT e AG, característicos sítios 5' e 3' dos íntrons. Estes alinhamentos geram forte evidência dos componentes das estruturas dos genes, muitas vezes definindo completamente a localização de cada éxon e cada íntron (Figura 4-4).

Procura ab initio *de genes*

A predição *ab initio*, por sua vez, depende tanto da informação de detectores de sinais quanto de conteúdo para delimitar a sequência gênica. Para tal, os algoritmos que se valem desta estratégia utilizam redes neurais, transformadas de Fourier e, mais comumente, modelos de Markov. Para realizar estas detecções, os algoritmos são treinados

com sequências conhecidas do genoma em questão. Por exemplo, a Figura 5-4 ilustra o grau de conservação dos nucleotídeos presentes na sequência de Kozak de *Drosophila melanogaster*, perfil este que pode ser utilizado na predição de novas sequências codificantes neste organismo. Outro exemplo pode ser observado no grau de conservação das regiões 5' e 3' provenientes de íntrons de genes humanos (Figura 6-4).

Dentre as limitações da predição *ab initio* está o fato de que, usualmente, o resultado obtido se refere às regiões codificantes, sem informações sobre regiões não traduzidas ou transcritos provenientes de *splicing* alternativo.

Assim, para sobrepujar estas limitações a combinação das duas estratégias parece ser a mais eficaz nos fluxos de trabalho utilizados para predição de genes em genomas sequenciados. Para tanto, alguns destes algoritmos são treinados com modelos gênicos já conhecidos, de organismos filogeneticamente próximos e, assim, provavelmente possuem uma estrutura gênica muito parecida com a do organismo que está em análise.

Anotação de regiões codificantes

O passo seguinte à identificação de sequências que possivelmente constituem genes é a sua anotação. A anotação manual foi bastante utilizada na análise dos primeiros genomas. Entretanto, devido à complexidade e ao alto número de sequências genômicas disponibilizadas a cada dia, há um consenso de que a anotação automática está se tornando indispensável.

A forma mais simples de anotação automática se dá pela análise de uma série de diferentes mecanismos de predição e delimitação de sequências gênicas e, então, utilização de um algoritmo de seleção, também denominado de *combiner*. Este algoritmo tem a função de selecionar a predição que melhor represente os modelos gênicos frente os algoritmos utilizados. Para tanto, os *combiners* estimam os tipos e as frequências de erros oriundos de cada programa de predição, escolhendo posteriormente as combinações de evidências que minimizam tais erros. Após as predições *ab initio* e baseados em evidência, alguns dos *combiners* devem ser treinados com sequências não previamente utilizadas nos programas de predições de genes.

Os *combiners* mais atuais utilizam técnicas que combinam evidências não estocásticas ponderadas (*nonstochastic weighted evidence*) que computam tanto o tipo quanto a abundância de uma evidência para o cálculo da sequência gênica consenso. Uma lista dos algoritmos mais utilizados para confecção de fluxos de trabalho para identificação de genes está disponível na Tabela 2-4.

A anotação da função de genes é um processo basicamente comparativo, sendo utilizados bancos de dados de proteínas, como o NCBI ou o UniProt (trEMBL + Swiss-Prot)

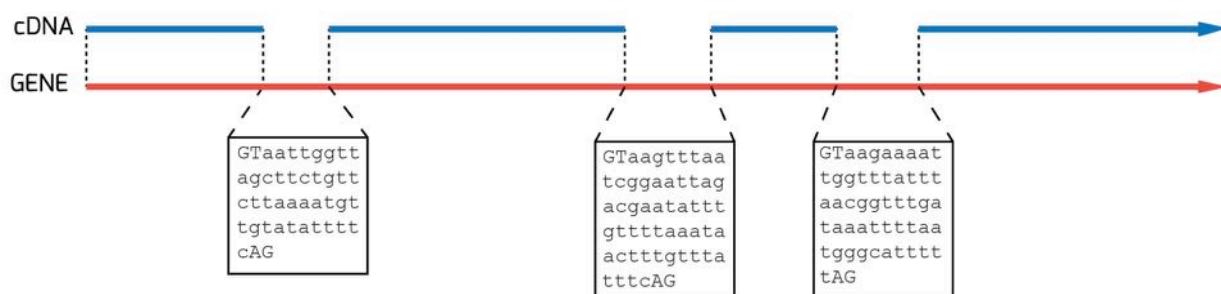



Figura 4-4: Identificação de genes baseada em evidência. Utilizando BLASTn com base em dados de transcritoma (cDNA, em azul), pode ser alcançada uma aproximação da sequência do gene (vermelho), inclusive permitindo a delimitação de éxons e íntrons. As regiões de identidade estão delimitadas por traços verticais. Com base na sequência de íntrons (quadros na porção inferior), é possível construir modelos para sua predição. Modelo construído com base no gene F10E9.5 de *Caenorhabditis elegans* (código de acesso NCBI NC_003281).

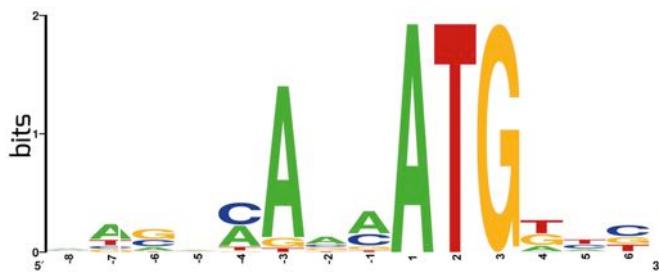



Figura 5-4: Padrão de conservação de nucleotídeos da sequência de Kozak, baseado no alinhamento de 30 sequências de cDNA obtidas de *D. melanogaster* e analisados junto ao servidor WebLogo. A medida de conservação é refletida pela altura da base. Os números abaixo representam o códon de início de tradução (1 a 3), o segundo códon do mRNA (4 a 6) e a região a montante (-8 a -1).

Figura 6-4: Padrão de conservação de nucleotídeos nas regiões 5' (painel superior) e 3' (painel inferior) de íntrons humanos. Resultado obtido pelo alinhamento de 100 sequências intrônicas e analisados junto ao servidor WebLogo. A medida de conservação é refletida pela altura da base. Os números abaixo de cada esquema indicam o início e o fim do íntron (0 e 1 no esquema superior; -2 e -1 no esquema inferior), assim como as regiões adjacentes.

ou de domínios proteicos (PFAM, NCBI CDD, Interpro). Uma das vantagens da utilização do Swiss-Prot como banco de dados para identificação dos produtos gênicos se refere ao fato deste ser um banco de dados manualmente curado, ou seja, inspecionado contra possíveis erros decorrentes da anotação automática. Com base nestas análises, quatro grupos distintos de anotações podem ser realizadas:

i) a existência de um ortólogo direto previamente caracterizado, revelado por BLAST, gerará a anotação com base no nome do ortólogo;

ii) a inexistência de um ortólogo direto, mas a presença de um domínio proteico conservado, revelado por análises em PFAM ou Interpro, gerará a anotação "*domain containing protein*" ou proteína contendo o domínio;

iii) a inexistência de ortólogos diretos previamente caracterizados ou domínios conservados confere as anotações proteína predita (*predicted protein*) ou proteína hipotética (*hypothetical protein*);

iv) quando um gene codificante de proteína hipotética possui ortólogos diretos, eles são denominados codificadores de proteína hipotética conservada (*conserved hypothetical protein*).

Outro passo na anotação da função de genes se refere à predição da localização da proteína codificada por este gene. Por exemplo, se uma proteína possui muitas regiões hidrofóbicas, compatíveis com sua inserção em membrana, possivelmente esta será uma proteína integral de membrana. Adicionalmente, proteínas secretadas ou endereçadas a alguma organela geralmente apresentam uma sequência sinal.

Diversas ferramentas estão disponíveis para localização de domínios transmembrana (TMHMM, TMPred, HMMTOp), baseando-se em métodos estatísticos para aferição da presença destes domínios. Métodos mais robustos para determinar a localização celular de um produto gênico foram desenvolvidos e se baseiam em uma diversidade de métodos estatísticos, geralmente treinados com sequências proteicas conhecidamente pertencentes a algum sub-compartimento celular (Tabela 3-4). De uma maneira geral, todas estas ferramentas são utilizadas na constru-

Tabela 2-4: Principais algoritmos utilizados na predição de genes e a sua funcionalidade.

Algoritmo	Descrição	Aplicação
Predições *ab initio* e baseados em evidência		
Augustus	Aceita evidências baseadas em transcriptomas e banco de dados de proteínas	Eucariotos
FGNESH	Arquivos para treino derivados de análise do fabricante	Eucariotos
fgenesB	Predição de genes e operons em bactérias baseadas em padrões e cadeias de Markov	Procariotos
Genemark	Arquitetura de busca baseada em *self-training*	Procariotos e eucariotos
Twinscan	Extensão do algoritmo Genscan que utiliza homologia entre dois genomas para guiar a predição de genes	Eucariotos
GenomeScan	Extensão do algoritmo Genscan que utiliza BLASTx para guiar a predição de genes	Eucariotos
Glimmer	Utiliza modelos de Markov interpolados	Procariotos
Combiners		
Evidence Modeler	Tem como resultado um modelo gênico pela combinação de evidências obtidas a partir de alinhamento de dados transcriptômicos e proteômicos com predições *ab initio*	Eucariotos
Evigan	Algoritmo de evidências probabilísticas que usa redes Bayesianas para pontuar e integrar predições *ab initio* e baseadas em evidência para produzir modelos gênicos.	Eucariotos

ção de fluxos de trabalho que integram diferentes ferramentas para analisar o resultado da predição de cada gene, conferindo uma anotação geral (Figura 7-4).

4.5. Identificação/anotação RNAnc

Considerando o dogma central da biologia molecular, no processo de síntese proteica (tradução) há a participação direta de pelo menos três classes distintas de RNAs:

i) o RNA mensageiro, que servirá de molde para síntese da proteína;

ii) o RNA ribossômico que, como indica o nome, é um componente estrutural e funcional dos ribossomos;

iii) o RNA transportador, que funciona como adaptador, carreando aminoácidos para serem incorporados na cadeia nascente da proteína durante o processo de tradução.

A anotação de genes de RNAs não codificantes - RNAnc (RNAt, RNAr, dentre outros) ainda não apresenta um grande número de programas quando comparada às estratégias disponíveis para anotação de genes codificantes de proteínas. Isto se deve, principalmente, à grande heterogeneidade e à pequena conservação dos RNAnc quando comparados a sequências de proteínas. Ao contrário de genes codificantes de proteínas, RNAnc geralmente não apresentam conservação de sequência $1^{ária}$, dificultando a detecção destes genes.

Um dos mecanismos mais utilizados na busca de RNAt em genomas é o tRNAscan-SE. Este algoritmo se baseia em uma série de cálculos estatísticos que avaliam, entre outros parâmetros, o potencial local para formação das estruturas $2^{árias}$ típicas de tRNAs em forma de trevo, assim como a presença de bases invariantes que definem regiões conservadas presentes nos promotores destes genes. Outro mecanismo de busca de RNAts se refere ao algoritmo ARAGORN. A

Tabela 3-4: Principais algoritmos utilizados na predição da localização celular de proteínas.

Algoritmo	Descrição	Aplicação
BaCelLo	Com base na composição de aminoácidos e sequências de treino, prediz em 5 localizações (secretada, citoplasmática, nuclear, mitocondrial e cloroplástica)	Plantas, animais e fungos
LOCtree	Com base na sequência N-terminal, prediz a localização em secretada, citoplasmática, nuclear, mitocondrial, cloroplástica e organelar.	Eucariotos e procariotos
TARGETp	Com base na sequência N-terminal, prediz a localização como secretada, mitocondrial e cloroplástica, dentre outras.	Eucariotos e procariotos
Wolf PSORT	Com base na sequência N-terminal e regras empíricas, classifica o endereçamento em cloroplástico, citosólico, citosesqueleto, retículo endoplasmático, extracelular, golgi, lisossômico, mitocondrial, nuclear, peroxissomal, membrana plasmática e membrana vacuolar. Permite localização múltipla.	Animais, fungos e plantas
Cell-PLoc	Permite realizar a localização de proteínas em mais de 25 diferentes locais, baseados em treino com sequências cuja proteína tem localização conhecida.	Eucariotos, procariotos e vírus

estratégia deste programa para a procura de tRNAs em sequências nucleotídicas se baseia em algoritmos heurísticos para a predição da estrutura do tRNA baseada na homologia com sequências conservadas, assim como a potencialidade de formar estruturas $2^{\text{árias}}$ típicas do tRNA. Por fim, o tRNAfinder se baseia em cálculos para detecção da estrutura $2^{\text{ária}}$ do RNA predito para identificar genes de tRNA.

Já a predição de RNArs é baseada em conservação de sequências. Ao passo que organismos procarióticos possuem geralmente três moléculas de RNAr (23S, 16S e 5S) completamente maduras e funcionais, eucariotos possuem quatro (28S, 18S, 5.8S e 5S). Cada uma destas sequências apresenta grande grau de conservação com os ortólogos de diferentes organismos. Desta forma, ferramentas baseadas em Modelos Ocultos de Markov, como o RNAmmer, foram construídas para delineamento dos genes responsáveis pelos RNArs. Adicionalmente, um grande banco de dados com famílias de RNA foi construído, e a cada ano novas adições de sequências de RNAs são feitas ao RFam. Estas famílias podem ser classificadas em três grandes grupos:

i) RNAs não codificantes (RNAnc);

ii) elementos estruturais regulatórios em *cis*, característicos de alguns RNAm que desempenham função de regulação da expressão gênica principalmente por meio da formação de estruturas $2^{\text{árias}}$;

iii) RNAs que podem sofrer o processo de *auto-splicing*.

Cada uma destas famílias é representada por alinhamentos múltiplos, consensos de estruturas $2^{\text{árias}}$ e modelos de covariância. Por meio de comparação de sequências com os consensos obtidos para os modelos de cada família, é possível identificar genes responsáveis pelos rRNAs, tais como os snoRNAs, que são componentes do spliceossomo. Existe ainda, contudo, uma grande gama de outros RNAnc que não apresentam grau de conservação necessário para formar uma família.

Identificação de pequenos RNAs

O termo "pequeno RNA" é, conceitualmente, muito vago e acaba englobando diferentes classes destes, como microRNAs, siRNAs, TAS-siRNAs, tRFs, entre outras. Contudo, existem características dos pequenos RNAs que podem ser utilizadas para identifi-

car as classes distintas: não codificam proteínas (apesar de alguns serem originados de regiões codificadoras), possuem tamanho variando entre poucas dezenas de nucleotídeos, suas rotas de biogênese e seus papéis funcionais.

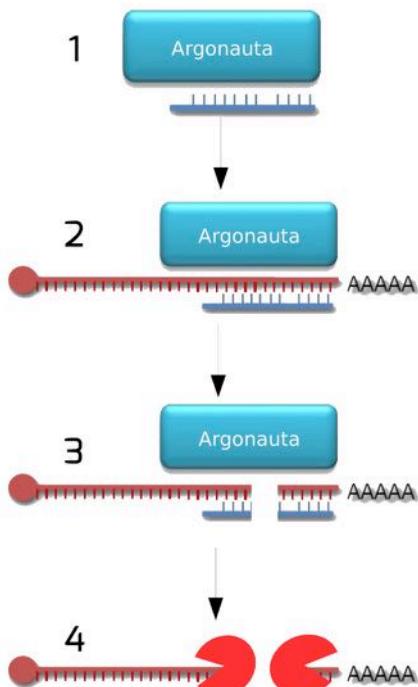
Os pequenos RNAs fazem parte de um grupo de pequenas moléculas, sendo conhecidos há décadas, e inicial e erroneamente creditados como produtos de degradação de RNA, não possuindo um papel biológico específico. Com a identificação do fenômeno de silenciamento gênico (RNAi) foi observado que pequenos RNAs poderiam, de fato, desempenhar um papel funcional, regulando a expressão gênica em vários níveis. Devido ao papel de forte regulador da expressão gênica, muita atenção tem sido dada aos pequenos RNAs, com um número crescente de trabalhos sendo feitos relacionando estes com patologias e controlando processos básicos do desenvolvimento.

Figura 7-4: Um fluxo de trabalho genérico para anotação de genes.

O RNAi, algumas vezes denominado de "silenciamento gênico", é um mecanismo que induz a diminuição da expressão gênica de um transcrito alvo através da clivagem do transcrito alvo e sua posterior degradação, ou através da repressão da maquinaria de tradução. Estes mecanismos são denominados também de Silenciamento Gênico Pós-Transcricional (PTGS – no inglês) (Figura 8-4). Existem adicionalmente alguns pequenos RNAs que induzem silenciamento gênico em nível transcricional, ligando-se em regiões de DNA, impedindo sua transcrição. Este mecanismo é denominado de Silenciamento Gênico Transcricional (TGS – no inglês).

As metodologias de sequenciamento de alta eficiência tem auxiliado de maneira contundente na caracterização de pequenos RNAs, sendo que variações de protocolos também possibilitaram validar alvos (técnica de degradoma) e identificar pequenos RNAs associados com proteínas específicas (sequenciamento de ácidos nucleicos associados a proteínas imunoprecipitadas).

Existe uma grande diversidade de pequenos RNAs em células eucarióticas, sendo os principais listados na Tabela 4-4. Dentre estas, os microRNAs são a classe de pequenos RNAs melhor descrita. Caracterizam-se por serem transcritos a partir de genes MIR, geralmente intergênicos, por uma RNA polimerase II, resultando em um pri-miRNA, o qual recebe um 5'-CAP e um 3'-poli-A. Este pri-miRNA é processado por um complexo proteico, denominado D-*body*, o qual é orquestrado por uma enzima classicamente denominada DICER ou DROSHA (RNAses classe III), resultando na liberação do pré-miRNA. Este apresenta estrutura em forma de grampo devido à alta complementaridade que suas extremidades 5' e 3' possuem. O pré-miRNA é

novamente processado por uma enzima DICER, liberando o microRNA maduro, dupla-fita, de aproximadamente 20 nucleotídeos de comprimento, o qual é reconhecido por uma enzima ARGONAUTA e direcionado ao PTGS (Figura 9-4).

Outra classe bastante estudada se refere aos siRNA (*small interfering* RNAs), os quais tem a biogênese bastante variada, podendo ser derivados de regiões de sobreposição de genes em orientação inversa natsiRNAs (*natural anti-sense small interfering* RNAs). A transcrição de ambos transcritos resulta em uma região de dupla-fita complementar, a qual é reconhecida por uma enzima DICER que cliva o natsiRNA, resultando na sua forma madura (aproximandamente 24 nt).

Existem também os tasiRNA (*trans-acting small interfering* RNAs), derivados do processamento do transcrito alvo de um microRNAs. Para a síntese de tasiRNA, é necessário uma RNA polimerase dependente de RNA, a qual utiliza o microRNA como iniciador da transcrição e a sequência transcrito alvo como molde. O longo RNA dupla-fita resultante é reconhecido também por uma enzima DICER, a qual cliva o tasiRNA, resultando na sua forma madura (aproximadamente 20 nt).

Os siRNAs são reconhecidos por enzimas argonautas e podem tanto induzir o silenciamento gênico por PTGS, mas também o remodelamento de cromatina, controlando a expressão gênica em nível trancricional (TGS). A interação entre microRNAs e transcrito alvo é a melhor caracterizada, não sendo necessário uma complementariedade perfeita entre o microRNA e transcrito alvo, apesar disto ser mais comum em plantas. Em animais existe uma região de maior complementariedade denominada *seed* a qual se localiza entre a 2ª e 7ª bases no microRNA, e está relacionada à especificidade do microRNA com seu transcrito alvo. Outra característica é o fato de ha-

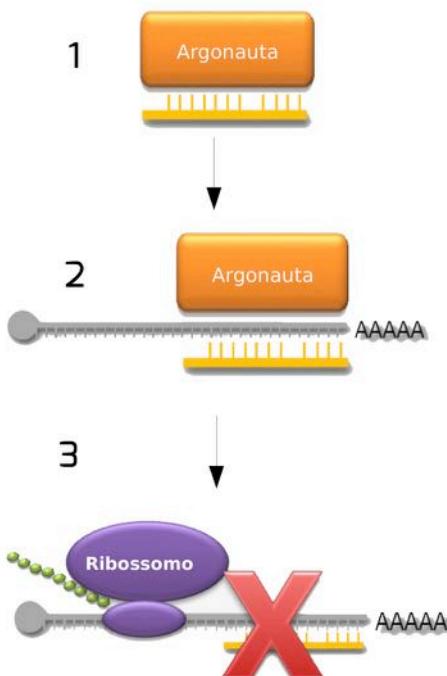



Figura 8-4: Mecanismo de PTGS. A) clivagem: 1, uma proteína argonauta reconhece uma fita do pequeno RNA; 2, O microRNA associado com uma argonauta reconhece um transcrito alvo; 3, ocorre a clivagem do transcrito alvo na posição medial do microRNA; 4, degradação do transcrito alvo clivado por nucleases. B) repressão da tradução: 1, uma proteína argonauta reconhece uma fita do pequeno RNA; 2, o microRNA associado com uma argonauta reconhece um transcrito alvo; 3, ocorre repressão da maquinaria de tradução.

Tabela 4-4: Principais classes de pequenos RNAs com função regulatória.

Classe	Tamanho (nt)	Função biológica	Mecanismo de ação	Origem	Organismos
microRNA ou miRNA	21-24	PTGS	Clivagem e repressão da maquinaria de tradução	Intergênica e íntrons	Plantas, animais, fungos e vírus
siRNA	21-24	PTGS, TGS	Clivagem, repressão da maquinaria de tradução e metilação de DNA	Intergênica, éxons e íntrons	Plantas, animais, fungos e vírus
tasiRNA	21-22	PTGS	Clivagem	Transcritos alvo de microRNAs	Plantas, animais e fungos
natsiRNA	21-22	PTGS	Clivagem	Transcritos convergentes parcialmente sobrepostos	Plantas

ver pareamento guanina – uracila (G-U), também denominado de *wobble* entre o transcrito alvo e o microRNA (Figura 9-4).

Existem dois desafios principais no emprego da bioinformática a pequenos RNAs. O primeiro é relativo à identificação da região, ou precursor, que dá origem ao pequeno RNA. O segundo envolve a identificação dos genes alvos regulados por estes. As metodologias de identificação da região que resulta no pequeno RNA variam com a classe de pequenos RNAs e estão intimamente relacionadas às suas biogêneses.

Os microRNAs são a classe melhor caracterizada, de forma que há uma maior disponibilidade de ferramentas para identificação destes, como os algoritmos miRTools, miRDeep, miRExpress, miRAnalyser e miRCat. A funcionalidade geral destes programas se baseia na análise de *reads* de sequenciamento de bibliotecas de pequenos RNAs e na delimitação das regiões de ancoramento com o genoma. Com base no conjunto de sequências ancoradas, são realizados cálculos para avaliação da estabilidade da possível estrutura em forma de grampo gerado pelo transcrito.

Para as demais classes, não existe uma metodologia padrão, sendo que variações da ferramenta BLAST são geralmente utilizadas. Para a identificar siRNAs, por exemplo, pode-se empregar a ferramenta SiLoCo. Mas é bastante comum laboratórios que pesquisam pequenos RNAs desenvolverem suas próprias ferramentas.

Já os programas de predição de alvos de microRNAs e siRNAs podem ser baseadas em ferramentas como o BLAST, procurando regiões complementares ao pequeno RNA. O problema é que esta técnica gera um número muito grande de falsos-positivos. Com isso, algumas ferramentas começaram a utilizar outros aspectos envolvidos na interação entre pequenos RNAs e transcritos alvos, tais como características energéticas, a presença da região *seed* (em humanos), o pareamento perfeito entre 10-11 pares de base do microRNA (válido somente para PTGS, por clivagem) e a conservação de microRNAs e transcritos alvo em organismos diferentes.

Mesmo assumindo estas regras, existem muitas interações entre microRNA e transcrito alvo que são excluídas, e muitas falsas que são incluídas, fazendo como que seja necessário a validação experimental desta interação. Especialmente para organismos modelo, existem bancos de dados próprios que disponibilizam, baseados em ferramentas de predição, os possíveis alvos para um determinado miRNA. Um importante banco de dados é o microRNA.org, cujas predições foram realizadas pelo algoritmo miRanda.

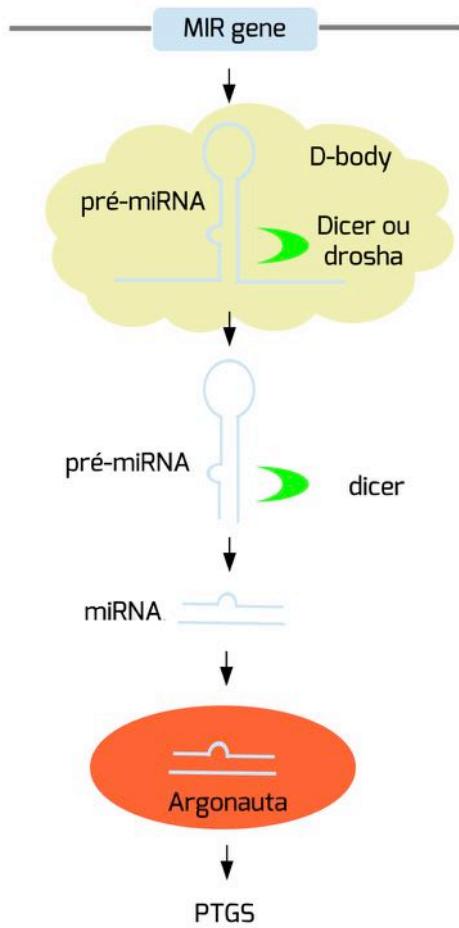



Figura 9-4: Modelo simplificado da biogênese de microRNAs. A partir de um gene MIR, um pré-miRNA é transcrito e processado num D-*body*, por uma enzima DICER, liberando o pré-miRNA, o qual é processado novamente por uma enzima DICER, liberando a forma madura do miRNA. Este é reconhecido por uma enzima argonauta e direcionado ao transcrito alvo, induzindo o silenciamento gênico.

4.6. Conceitos-chave

Anotação funcional: conjunto de abordagens que predizem a função e classificam uma proteína codificada por um genoma.

Contig: conjunto de segmentos de DNA com sobreposição de sequência que, conjuntamente, representam uma sequência consenso de DNA

Detectores de conteúdo: sistemas para delimitação de regiões codificantes baseados na classificação da sequência em codificante ou não codificantes, baseada em cálculos estatísticos ou em conservação de sequência. Compreendem detectores extrínsecos e intrínsecos.

Detectores de sinais: sistemas para delimitação de regiões codificantes baseados em caracteres funcionais de genes, como elementos canônicos necessários à transcrição ou tradução.

N50: índice associado à qualidade de montagem de um sequenciamento. Um valor de N50 igual a N significa que 50% dos *reads* estão montados em um *contig* de tamanho N ou maior.

ORF: *open reading frame* ou fase aberta de leitura. Refere-se a toda sequência nucleotídica delimitada por um códon de início e um códon de término de tradução.

Predição baseada em evidência: identificação de sequências codificantes baseada em experimentos prévios, como transcriptomas.

Predição *ab initio*: identificação de sequências codificantes baseada unicamente em cálculos estatísticos.

Reads: resultado obtido do sequenciamento de um determinado clone ou fragmento de DNA/cDNA.

Sequenciamento por *Shotgun*: metodologia de sequenciamento caracterizado por fragmentação aleatória de um grande segmento de DNA, determinação individual da sequência de cada um dos fragmentos e agrupamento dos *reads* obtidos em *contigs*.

Sinais transcricionais: sequências conservadas associadas ao processo de transcrição, como por exemplo TATA box, Sítios de clivagem e poliadenilação, etc.

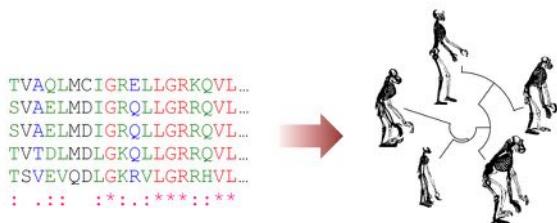
Sinais traducionais: sequências conservadas associadas ao processo de tradução, como a sequência de Kozak, códon de início de

tradução, sítio de ligação de ribossomo, etc.

Transcriptoma: sequenciamento e avaliação geral de transcritos de uma célula/tecido com o intuito de descrever os RNAs presentes naquele momento. Além de trazer informações sobre a situação fisiológica daquele conjunto de células, permite construir modelos para procura de genes baseados em evidência.

4.7. Leitura recomendada

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. ***Nat. Methods***, 8, 469-477, 2011.

RICHARDSON, E. J.; WATSON, M. The automatic annotation of prokaryotic genomes. ***Brief. Bioinform***., 14, 36-45, 2013.

SLEATOR, R. D. An overview of the current status of eukaryotic prediction strategies. ***Gene***, 461, 1-10, 2010.

WILLIANSON, V. et al. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. ***Brief Bioinform***., 14, 36-45, 2013.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. ***Nat. Rev. Genet***., 13, 329-342, 2012.

5. Filogenia Molecular

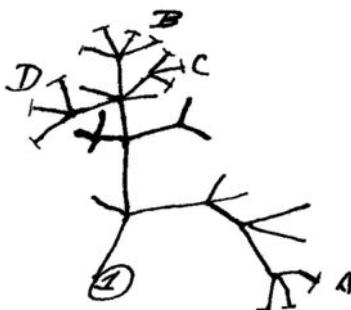
Rodrigo Ligabue Braun
Dennis Maletich Junqueira
Hugo Verli

Estabelecimento de relações evolutivas a partir de sequências de aminoácidos ou nucleotídeos.

5.1. Introdução

Desde seus primórdios, a humanidade se mostrou inclinada a organizar e classificar o mundo à sua volta com o objetivo de facilitar o entendimento e a comunicação. Em relação ao mundo natural, diferentes sistemas foram empregados para compor métodos de organização e classificar os organismos, utilizando critérios naturais ou artificiais.

Um dos sistemas de maior influência no período pré-Darwiniano foi a Escala Natural de Platão. Neste sistema, do fogo ao ser humano, diferentes níveis eram organizados à maneira de uma escada. A ideia de ascensão estava associada à perfeição, representada em sua forma plena pelo homem. O sistema classificatório de Lineu, por sua vez, se baseava em características visíveis, arbitrariamente selecionadas para classificar os seres vivos (por exemplo, número de patas ou de pétalas), sendo o ser humano o organismo do topo da cadeia. Sistemas como este são considerados sistemas artificiais, pois estão sujeitos à tendência de seu autor em considerar um caractere em detrimento de outro(s), conforme sua vontade ou necessidade. Entretanto, como o próprio Lineu reconheceu, tais sistemas foram absolutamente necessários para a fase inicial (descritiva) da biologia, servindo de base para o sistema natural de classificação e para as hipóteses de similaridade que surgiriam a seguir.

Ao final do século XVIII e início do século XIX, surgem os sistemas naturais de classificação. Estes buscavam refletir sobre a ordem natural dos seres vivos através de poucas características intrínsecas, geralmente associadas à forma. No entanto, com o objetivo de tornar a classificação mais racional, tomaram lugar debates sobre a real necessidade de haver um sistema hierárquico de organização dos organismos. Opositores da ideia consideravam que a classificação era, muitas vezes, inadequada e desnecessária, e que não deveria ser um fim em si mesma, senão um método para o levantamento de novas perguntas à Biologia.

Em 1818, a introdução do conceito de homologia por E.G. Saint-Hillaire causa uma revolução nas ciências biológicas. Para ele e seus colegas, partes homólogas correspondiam às partes de animais diferentes com uma estrutura essencialmente semelhante, mesmo com forma ou função distintas. Por

exemplo, as asas de um morcego, as nadadeiras de uma baleia e os braços de um macaco, segundo esta lógica, são considerados órgãos homólogos e podem servir como critério para agrupar morcegos, baleias e macacos em um mesmo grupo. Assim, a homologia serviria como critério principal para uma classificação natural dos organismos.

A partir da famosa publicação de Darwin, "A Origem das Espécies", em 1859, a classificação dos organismos passou a ser não apenas natural, mas também a apresentar uma condição essencial de ancestralidade comum. Segundo este pensamento, os organismos são derivados uns dos outros, desde o surgimento da vida na terra. Darwin representou este padrão através de um esquema de ramificação, onde os galhos representam o tempo entre o organismo ancestral e o novo organismo, e os nós representam os próprios organismos. Mais tarde, esta viria a ser a primeira árvore filogenética utilizada para representar processos evolutivos.

A primeira árvore filogenética moderna (esboço de Darwin no manuscrito de A Origem das Espécies)

Com influência direta da teoria evolutiva de Darwin (e colaborações de Wallace e Lamarck), desenvolve-se a Taxonomia Evolutiva. Este sistema de classificação incorporou o vetor tempo (caráter temporal normalmente inferido por meio de fósseis) e, além disto, adicionou uma quantificação da divergência estrutural entre os grupos (a chamada distância patrística). Já em meados do século XX, inicia-se a Fenética (taxonomia numérica ou neodansoniana). Esta escola buscava incluir na classificação dos organismos o máximo possível de características, atribuindo-lhes o mesmo peso na tentativa de eliminar qualquer subjetividade ou arbitrariedade. Seu impacto, entretanto, foi limitado devido às dificuldades em traduzir os índices (valores) obtidos em informações relevantes do ponto de vista biológico (como a separação de espécies, por exemplo). Na mesma época, surge a Cladística (ou sistemática filogenética), liderada pelo entomólogo alemão Willi Hennig. Na proposta de Hennig (1950), organismos que compartilhassem características derivadas (apomórficas) poderiam ser considerados descendentes do organismo ancestral, na qual a característica em seu estado primitivo (ou plesiomórfico) passou para o estado derivado.

Desde a origem dos sistemas de classificação até a Cladística, os métodos baseavam-se essencialmente no fenótipo dos organismos, ou seja, em suas características físicas claramente discerníveis. Entretanto, com o advento dos métodos de sequenciamento, tanto protéico quanto genômico, cada vez mais os dados moleculares foram se tornando importantes nas análises evolutivas de ancestralidade. Neste sentido, a ciência passa de um ponto de vista macroscópico a um ponto de vista molecular de análise.

O método de sequenciamento de aminoácidos, iniciado por Sanger em 1954, abriu caminho para que proteínas de uma mesma classe, em diferentes organismos, pudessem ser comparadas quanto às suas origens evolutivas. Da mesma forma, ao decodificar a primeira longa sequência de DNA, em 1977, Sanger deu início à explosão do sequenciamento de ácidos nucleicos, permitindo a comparação de genes em larga escala. É importante destacar que as sequências moleculares podem tanto ser comparadas entre si, buscando conhecer a história evolutiva de um gene ou proteína (por exemplo, relações entre hemoglobinas de diferentes mamíferos), quanto podem ser associadas a outros dados na reconstrução da história evolutiva de organismos (por exemplo, associando as relações obtidas por comparação de DNA ribossomal de aves com datação de fósseis, buscando estabelecer relações de ancestralidade).

No entanto, ao lidar com sequências moleculares, diferentes questões podem surgir. Por exemplo, o conceito de gene é di-

nâmico e mudou muito desde sua primeira definição. Além disso, genes podem sofrer diferentes processos evolutivos que alteram sua estrutura e/ou função, como mutações e rearranjos, ou ainda duplicações e perdas de função. Esses fatores fazem com que a relação 1:1 entre gene e organismo seja perdida. Por exemplo, uma mesma leguminosa pode possuir duas cópias do gene para a proteína leghemoglobina (genes parálogos). Além disso, muitas sequências do genoma não chegam à etapa de tradução, podendo conter elementos regulatórios ou transponíveis. Tais variações aumentam a complexidade e dificultam a interpretação das relações de descendência.

5.2. Aplicações

Ao classificarmos os organismos, atribuímo-lhes uma história evolutiva. Essa história, entretanto, é frequentemente desconhecida. Sendo assim, é necessário inferir a sequência de mudanças que levaram ao surgimento de um novo organismo ou proteína. Contudo, existe apenas uma história verdadeira, que talvez jamais seja conhecida. Assim, ao empregarmos as técnicas filogenéticas, o objetivo é coletar e analisar dados capazes de fornecer a melhor estimativa para chegarmos à filogenia verdadeira. De certa forma, a obtenção de filogenias lembra a atuação de um historiador. Baseando-se em dados disponíveis no presente (tais como organismos vivos, fósseis e sequências moleculares), tenta-se obter uma imagem de como teria sido o passado.

Quando analisamos sequências de nucleotídeos ou aminoácidos para inferir uma filogenia, utilizamos informações derivadas das taxas evolutivas para determinar a sequência de eventos que levaram ao surgimento de novos organismos. A taxa de evolução molecular refere-se à velocidade na qual os organismos acumulam diferenças genéticas ao longo do tempo. Essa taxa é frequentemente definida pelo número de substituições por sítio (ou posição no alinhamento de sequências) por unidade de tempo e, portanto, são usadas para descrever a dinâmica das mudanças em uma linhagem ao longo de várias gerações.

As taxas evolutivas são empregadas quando se buscam estimativas temporais para datação de eventos evolutivos. Normalmente, se assume que as mudanças nas sequências se acumulam a uma taxa mais ou menos constante ao longo do tempo. Esse conceito é chamado de Hipótese do Relógio Molecular. Entretanto, é conhecido que as taxas evolutivas são dependentes de vários fatores, tais como o tempo de geração, o tamanho da população e do próprio metabolismo, o que normalmente viola o modelo estrito de relógio molecular. Com base nestas informações, diversos modelos foram propostos para lidar com desvios no comportamento temporal de diferentes linhagens moleculares e, hoje em dia, são referidos como relógios moleculares relaxados.

Atualmente, a inferência filogenética é um campo de pesquisa à parte das outras ciências. Tornou-se uma ferramenta complementar para diversas áreas e indispensável para outras. Apesar de ter sido idealizada para desvendar apenas as relações evolutivas entre organismos, atualmente a filogenética molecular é aplicada a problemas muito mais diversos que este. Com o advento do relógio molecular estrito, foi possível aplicar a estimativa de tempo às filogenias e datar surgimento de espécies, disseminação de organismos e, até mesmo, entender grandes eventos biológicos que ocorreram no passado. Com a abordagem relaxada do relógio molecular, iniciou-se a utilização de modelos de dinâmica populacional que comportam os eventos coletivos de grupos específicos. Ainda, com o avanço da capacidade de processamento computacional, vem sendo possível criar algoritmos capazes de reconstruir genomas ancestrais. Também a partir da filogenética molecular desenvolveu-se o campo da filogeografia. Segundo esta área do conhecimento, as filogenias podem ser utilizadas para verificar a distribuição geográfica de indivíduos. Neste contexto, outras técnicas, além das filogenias, são incorporadas às aná-

lises, incluindo a estruturação de genes, as análises de redes e as análises de haplótipos.

A filogenia molecular busca inferir a história evolutiva de organismos ou outras entidades biológicas (como proteínas e genes) a partir de sequências de ácidos nucleicos ou aminoácidos. Ao investigar as relações entre diferentes espécies, análises de genes ribossomais são comumente empregadas, pois independentemente da espécie ou do organismo, os indivíduos possuirão genes codificantes de RNA ribossômico. Em contrapartida, quando se busca compreender as relações entre diferentes enzimas de uma mesma família é necessário utilizar sequências de aminoácidos, e não de nucleotídeos. Em determinadas situações, o genoma completo pode ainda ser utilizado para inferir a filogenia. Este é o caso de diversos vírus, especialmente quando se busca compreender a origem de novas variantes ou a disseminação de uma cepa. O alvo de estudo (isto é, sequência de nucleotídeos ou aminoácidos, gene ou genoma) depende, exclusivamente, do objetivo da análise e é um dos principais fatores a ser definido primariamente pelo pesquisador.

Atualmente, as filogenias funcionam como importantes ferramentas para diferentes áreas do conhecimento, incluindo as áreas de evolução, genética, epidemiologia, microbiologia, virologia, parasitologia, botânica e zoologia, dentre outras. Adicionalmente, de maneira inédita, a inferência filogenética foi utilizada como evidência para a resolução de crime e principal prova durante um impasse internacional envolvendo diferentes países. Em resumo, dependendo do objetivo, os métodos de construção de filogenias (inferência filogenética) são a base para diversas áreas e importantes objetos para o avanço computacional na análise de dados biológicos.

5.3. Representação de árvores

A Filogenética (termo obtido por união dos termos gregos para tribo e origem) é a ciência que busca reconstruir a história evolutiva dos organismos, levando em conta as sequências de nucleotídeos ou aminoácidos. As hipóteses sobre a história evolutiva são o resultado dos estudos filogenéticos e se chamam Filogenia.

As filogenias ou árvores filogenéticas representam o contexto evolutivo dos organismos de forma gráfica. São formadas por nós (pontos) ligados por diversos ramos (linhas) (Figura 1-5). Os nós terminais, mais externos na filogenia, identificam os indivíduos, genes ou proteínas que foram amostrados e incluídos na análise filogenética. Geralmente representam o alvo de estudo do pesquisador e estão ligados aos nós mais internos na filogenia através de traços horizontais, chamados de ramos terminais (Figura 1-5).

Os nós internos, pelo contrário, representam indivíduos não amostrados. Eles identificam uma inferência evolutiva do ancestral comum mais recente dos ramos derivados daquele nó e se ligam a nós cada vez mais internos, através dos ramos internos. Por exemplo, na Figura 1-5, os grupos de nós terminais representados em verde possuem como ancestral comum o nó laranja, mais interno, enquanto os nós terminais azuis possuem como ancestral comum o nó lilás. Da mesma forma, o nó vermelho é a representação do indivíduo, gene ou proteína mais ancestral da filogenia que, através de processos evolutivos, deu origem aos nós laranja e lilás.

O tamanho dos ramos horizontais pode ter diferentes significados, dependendo do método para inferência da filogenia, conforme

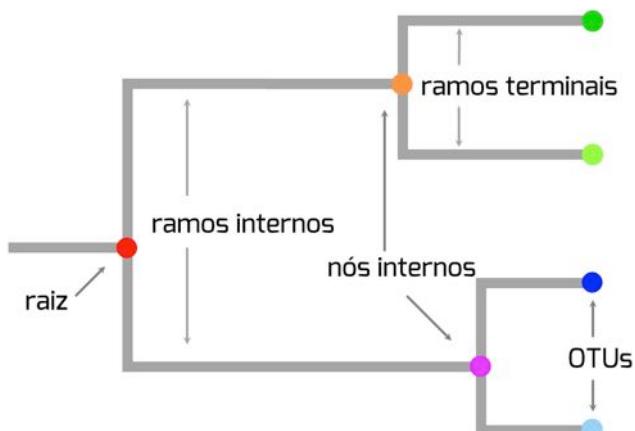



Figura 1-5: Nomenclatura associada a árvores filogenéticas.

veremos a seguir. No entanto, os ramos representados na vertical (Figura 1-5) não expressam qualquer significado, e seu tamanho não altera em nada a idéia filogenética. Como a análise pode ser feita em diferentes níveis, utilizando dados moleculares de genes, proteínas, indivíduos, espécies, gêneros, famílias, ou qualquer outro taxon, os nós terminais são amplamente denominados OTUs (*operational taxonomical units*), ou unidades taxonômicas operacionais (também chamados de folhas, Figura 2-5). A ordem e disposição exata das OTUs em uma filogenia é denominada topologia.

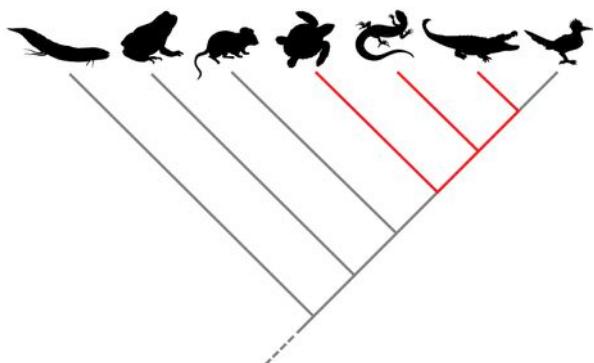



Figura 2-5: Árvore dicotômica dos grupos de vertebrados. As OTUs (nós terminais) estão representadas por ícones (peixes pulmonados, anfíbios, mamíferos, tartarugas, lagartos e serpentes, crocodilos e aves). Observe que o grupo dos répteis é parafilético (destacado em vermelho). O grupo seria considerado monofilético se incluísse as aves.

Além da forma gráfica, as árvores filogenéticas podem também ser descritas na forma textual. Em vez do diagrama com linhas e pontos, as relações evolutivas são representadas por notações com parênteses. A estrutura da árvore da Figura 2-5, por exemplo, pode ser descrita linearmente como (Peixes pulmonados, (Anfíbios, (Mamíferos, (Tartarugas, (Lagartos, (Crocodilos, Aves)))))) ou (Peixes pulmonados + (Anfíbios + (Mamíferos + (Tartarugas + (Lagartos + (Crocodilos + Aves)))))). Estas notações foram desenvolvidas para utilização computacional da informação filogenética. Algoritmos e programas que realizam análises moleculares necessitam da informação na forma textual e, quando necessário, fornecem a saída para o usuário na forma gráfica.

Partindo do princípio de derivação evolutiva, onde um organismo dá origem a outro (ou outros), podemos reconhecer dois principais processos na representação de filogenias: derivação dicotômica e derivação politômica. No primeiro caso, cada nó interno dá origem a apenas dois ramos. Para espécies, por exemplo, a ramificação de um ancestral comum em dois ramos evidencia o processo de especiação. No segundo caso, três ou mais ramos surgem de um mesmo nó interno.

Apesar de árvores dicotômicas serem mais comuns e normalmente esperadas, em alguns casos, como a dispersão explosiva do HIV e do HCV, árvores politômicas representam melhor o processo evolutivo. Casos como estes, onde um ancestral comum origina simultaneamente várias linhagens descendentes, são chamadas de politomias verdadeiras (*hard polytomies*). Por outro lado, as politomias falsas (*soft polytomies*) são casos onde a topologia não foi bem resolvida por não haver certeza do padrão de ancestralidade, tornando múltipla uma divisão que se esperaria ser formada por uma série de divisões dicotômicas.

Assim, ao agruparmos as OTUs segundo a sua ancestralidade, podemos reconhecer diferentes padrões: grupos monofiléticos, parafiléticos e polifiléticos (Figura 2-5). Os grupos monofiléticos incluem todos os membros descendentes de um único ancestral, assim como o próprio ancestral. Na Figura 2-5, por exemplo, as aves e os crocodilos são considerados um grupo monofilético, pois compartilham o mesmo ancestral comum. Da mesma forma, as aves, os crocodilos e os lagartos também podem ser considerados um grupo monofilético, pois se originaram de um mesmo ancestral. A análise das relações entre os grupos, neste caso, dependerá do objetivo do pesquisador. Adicionalmente, os grupos monofiléticos podem ser denominados clados por agruparem duas ou mais sequências que são descendentes de um mesmo ancestral (Figura 3-5a e b). A organização da topologia em que um clado está contido em outro é comumente chamada de clados aninhados ou clados embutidos (Figura 3-5c).

Os grupos parafiléticos, por sua vez, se

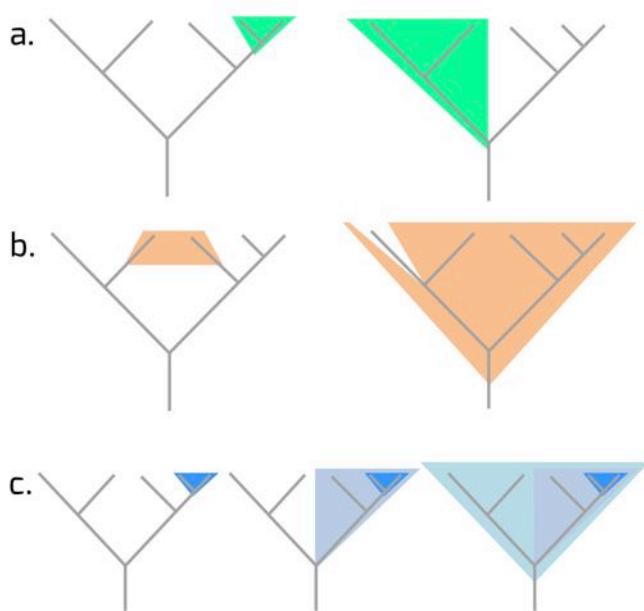



Figura 3-5: (a) Exemplos de clados destacados em verde. (b) Exemplos de organizações da topologia que não caracterizam a existência de um clado, destacados em laranja. (c) Diferentes níveis de clados que podem estar embutidos em um clado de maior ordem. Observe que os clados de diferentes ordens, quando embutidos, formam clados monofiléticos.

originam de um único ancestral, mas nem todos os organismos derivados deste ancestral fazem parte do grupo. Na Figura 2-5, os répteis são um grupo formado pelas tartarugas, lagartos e crocodilos, e seu ancestral comum está na base do ramo que dá origem às tartarugas. No entanto, este ancestral comum também deu origem às aves e, por isso, os répteis não podem ser considerados um grupo monofilético, mas um grupo parafilético.

Finalmente, os grupos polifiléticos provêm de dois ou mais ancestrais diferentes. Nestas relações se encontram OTUs que apresentam características comuns, mas que possuem diferentes ancestrais comuns. Por exemplo, a condição endotérmica (animais que mantém a sua temperatura corporal constante) é apenas apresentada por aves e mamíferos. Por este critério, poderíamos agrupar estes dois grandes grupos sem, no entanto, compartilharem o mesmo ancestral comum direto (Figura 2-5). A organização destes grupos permite descrever características resultantes de convergência evolutiva, pois uma mesma característica se desenvolveu independentemente em diferentes grupos.

Sabendo das relações evolutivas entre os táxons e da existência de ancestrais comuns, as árvores podem ser representadas de maneira a evidenciar o ancestral mais antigo (árvore com raiz ou enraizada), ou apenas destacar as relações evolutivas entre os táxons, sem destacar qual a OTU mais ancestral (árvore sem raiz ou não enraizada) (Figura 4-5).

A raiz da filogenia é a espécie ou sequência ancestral a todo o grupo que está sob análise. Quando presente, a raiz aplica uma direção temporal à árvore, permitindo observar o sentido das mudanças evolutivas da raiz (mais antigo) aos ramos terminais (mais modernos). Uma árvore não enraizada, pelo contrário, reflete apenas a topologia estabelecida entre as OTUs, sem indicar o ancestral do grupo. Árvores não enraizadas podem ser confusas, e sua interpretação requer mais cuidado devido à facilidade em cometer erros de análise (Figura 4-5).

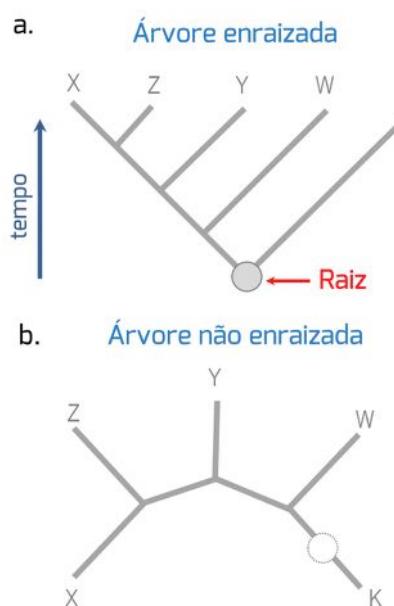



Figura 4-5: Comparação de árvores (a) enraizadas e (b) não enraizadas. No primeiro caso, é possível definir a direção das mudanças evolutivas, devido à presença do vetor tempo dado pela presença da raiz.

A identificação de uma raiz nas filogenias geralmente requer a inclusão de uma ou diversas OTUs que representem grupos externos. Os grupos externos devem ser ancestrais comuns das OTUs em estudo, já conhecidos, que indicarão caracteres presentes em organismos mais próximos aos ancestrais, provendo um direcionamento para a interpretação dos processos evolutivos. Para o caso do estudo de HIV, por exemplo, é comum que os vírus da imunodeficiência de símios (SIV) sejam utilizados como grupo externo nas filogenias, pois sabidamente estes vírus deram origem ao HIV.

A adição de grupos externos aumenta o número de topologias diferentes que uma filogenia pode assumir. O número de árvores possíveis varia com o número de OTUs e com a presença ou ausência de raiz. Para mais de duas OTUs, a quantidade de possíveis árvores com raiz é sempre maior que o número de árvores sem raiz. A possibilidade de inferência de diferentes topologias para os mesmos dados moleculares ressalta a extrema variabilidade de cenários possíveis na busca do verdadeiro evento evolutivo. É importante também ressaltar que, assim como a complexidade, o tempo computacional envolvido na construção das filogenias aumenta exponencialmente com o aumento de OTUs.

Em relação à topologia das árvores, a inversão de ramos derivados de um mesmo nó não altera a relação evolutiva apresentada pela árvore (Figura 5-5). Nesse sentido, a árvore filogenética pode ser comparada a um móbile: cada peça suspensa é livre para girar em seu eixo, ficando mais próxima ou mais distante espacialmente das outras peças, sem alterar a estrutura geral do objeto. Independentemente da posição destas OTUs, após o giro dos ramos, o mesmo ancestral comum será identificado e, por isso, não há qualquer alteração no significado da filogenia.

Figura 5-5: A porção terminal da árvore dos vertebrados (representada na Figura 2-5) foi rearranjada de diferentes maneiras (as setas indicam o ponto de rotação). Conforme a analogia de um móbile, todas elas representam a mesma relação evolutiva.

Quanto à nomenclatura de árvores filogenéticas, diferentes termos são empregados, tais como cladogramas, filogramas e dendrogramas (Figura 6-5). Um cladograma é uma árvore simples, que retrata as relações entre os nós terminais. Pelo contrário, uma árvore aditiva (árvore métrica ou filograma) apresenta informações adicionais, pois o comprimento dos ramos é proporcional a algum atributo, como quantidade de mudança. Por sua vez, uma árvore ultramétrica (ou dendrograma) constitui um tipo especial de filogenia devido aos seus ramos serem equidistantes da raiz. Os dendrogramas podem, desta forma, retratar o tempo evolutivo. É importante ressaltar que alguns autores denominam qualquer filogenia como cladograma, o que pode ser confuso.

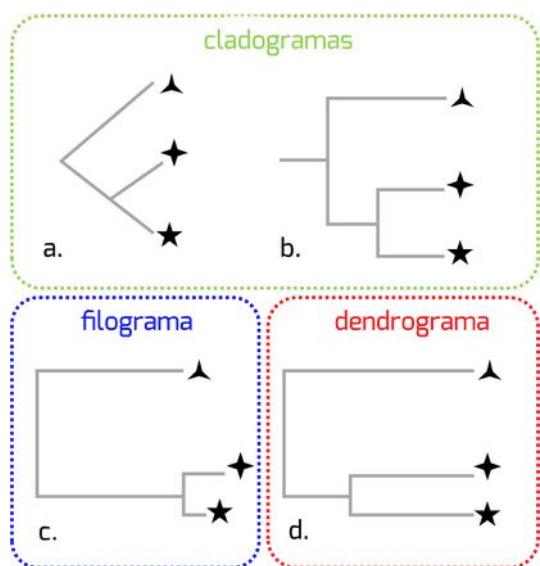



Figura 6-5: Nomenclatura de árvores filogenéticas. Observe que os cladogramas a e b são equivalentes, mas o filograma c e o dendrograma d não o são.

O tipo de dado molecular a ser empregado nas análises também deve ser levado em conta. Sequências de aminoácidos são mais conservadas que sequências de ácidos nucleotídeos em decorrência da degeneração do código genético. São, portanto, úteis em análises de produtos de genes ou espécies que visam entender fenômenos que aconteceram há amplos períodos de tempo evolutivo. Além disso, por formarem um conjunto de pelo menos 20 membros (contra quatro membros presentes em DNA ou RNA), sua variação pode ser mais significativa.

A despeito desta diferença no volume de informação, com a popularização do sequenciamento de ácidos nucleicos, especialmente DNA, sequências de nucleotídeos passaram a ser as mais empregadas em estudos de filogenia. Ácidos nucleicos são mais propensos a alterações, podendo sofrer transições (quando ocorre a troca de uma purina por outra purina, ou de uma pirimidina por outra pirimidina) e transversões (quando ocorre a troca de uma purina por uma pirimidina ou vice-versa), além de inserções ou deleções de pares de base que interferem no quadro de leitura. Essa variabilidade pode ser interessante no estudo de eventos mais recentes do ponto de vista evolutivo.

É preciso, assim, conhecer o caso de estudo e o tipo de pergunta que se busca responder com cada filogenia. Ao lidarmos com genes de diferentes espécies, por exemplo, é importante saber da existência e disposição de íntrons, da necessidade de lidar com o gene inteiro ou apenas parte dele ou da necessidade de incluir regiões regulatórias para a análise.

Um exemplo recente da aplicação de análises filogenéticas está no caso da identificação da origem da linhagem do vírus influenza H1N1, envolvido no surto de gripe de 2009. Para tanto, Smith e colaboradores empregaram genomas completos de influenza isolados de diferentes localidades e hospedeiros, e construíram árvores filogenéticas para cada uma das oito regiões do genoma buscando identificar a fonte de cada rearranjo presente no vírus envolvido no surto. Por meio das árvores obtidas, foi possível rastrear a contribuição genética dos vírus isolados de aves, suínos e humanos (Figura 7-5). Assim, o emprego da filogenia neste trabalho permitiu não apenas caracterizar o vírus do ponto de vista molecular, como também reconstruir a história evolutiva do agente etiológico de uma pandemia.

5.4. Distância genética

A formulação de modelos evolutivos é uma maneira de descrever matematicamente os processos que moldam as mudanças nas sequências de nucleotídeos ou aminoácidos dos organismos ao longo do tempo. Do ponto de vista molecular, estas mudanças podem ser resultado de diferentes forças evolutivas que reorganizam a sequência e a própria estrutura dos genes.

Um modelo geral para descrever de maneira eficaz estas alterações evolutivas deveria considerar os processos de substituição, inserção, deleção e duplicação, bem como ocorrência de transposição ou até mesmo de retrotransposição. Contudo, apesar de estes fenômenos serem claros agentes na modelagem dos genomas, matematicamente

vírus H1N1 aviário
PB2,PA
H1N1
M
N1
vírus H1N1 suíno clássico
H1,N1
NP,M,NS
surto humano com gripe suína
infecção humana esporádica com vírus recombinante suíno
PB2,PB1, PA, H1,NP,NS
surto humano com vírus H1N1 suíno
H3,N2,PB1
vírus H3N2 humano sazonal
aviário
suíno
humano
1979
1992
1997
1998
2009

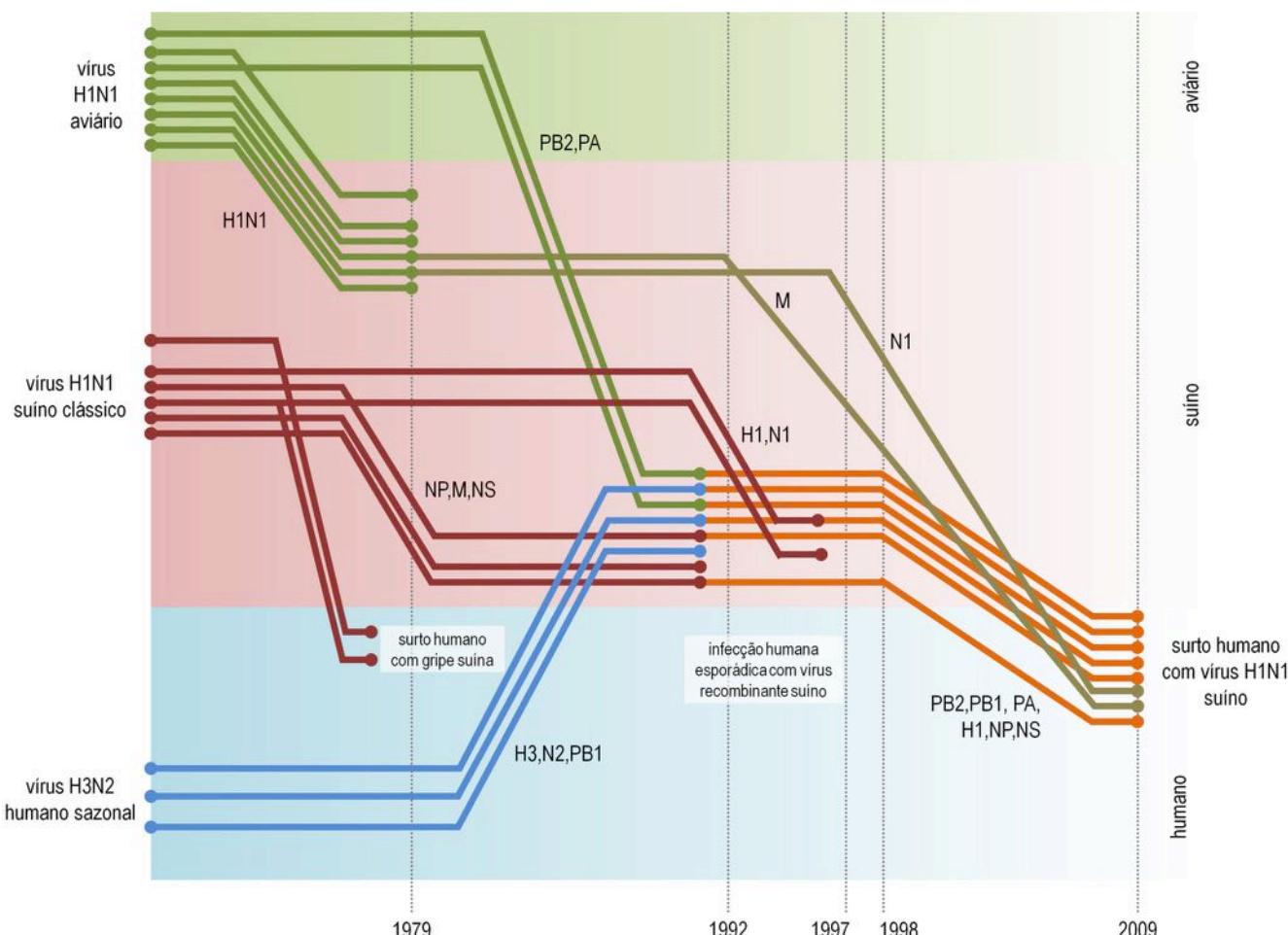



Figura 7-5: Representação esquemática das recombinações que originaram o vírus Influenza envolvido no surto de gripe suína em 2009. Diferentes linhas representam diferentes regiões do genoma do vírus. Observe a interação entre vírus de origens aviária, suína e humana em eventos que datam, pelo menos, desde 1990. Os eventos de recombinação e as análises temporais foram baseadas em análises filogenéticas (Adaptado de Smith e colaboradores, *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic*. *Nature*, *459*, 1122-1125, **2009**).

ainda não é factível colocá-los como componentes de modelos que expliquem inteiramente o processo evolutivo.

Assim, devido à grande relevância dos mecanismos de substituição para a evolução dos genomas em diferentes organismos e da disponibilidade de modelos de probabilidade estatística que expliquem este processo, as trocas têm sido o principal alvo para o desenvolvimento de modelos matemáticos e compõem a base de diversos métodos de inferência filogenética.

Após a divergência de duas sequências a partir de seu ancestral comum, de forma dicotômica, fenômenos evolutivos garantirão as mudanças nas sequências de nucleotídeos de forma independente (Figura 8-5). Uma medida tradicional para expressar o número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência é chamada de distância genética. Esta informação é uma medida quantitativa da dissimilaridade genética entre diferentes OTUs, e permite estabelecer uma estimativa relativa da quantidade de mudanças que ocorreram desde a divergência.

A distância é também um importante conceito na construção de filogenias, pois está diretamente relacionada com a relação evolutiva entre duas OTUs: uma menor distância

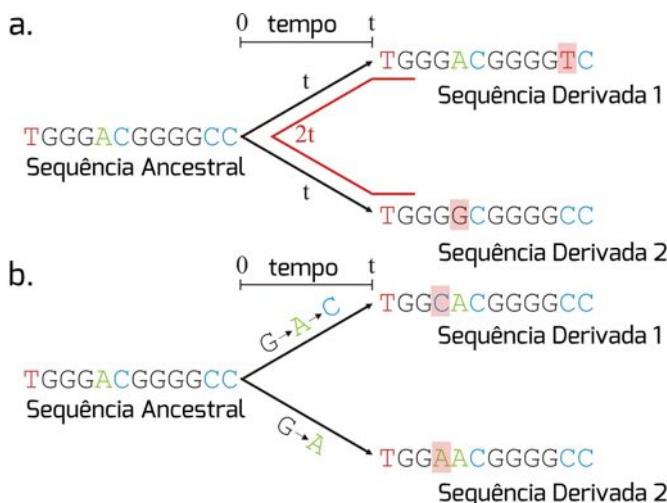



Figura 8-5: Após a divergência de dois organismos a partir de seu ancestral comum, seus genomas acumularão diferenças independentemente. (a) A medida da dissimilaridade genética entre duas sequências homólogas ao longo do tempo é chamada de distância genética, e a relação temporal entre duas sequências divergentes é dada por $2t$. (b) A ocorrência de múltiplas substituições ao longo do tempo na divergência de sequências homólogas pode mascarar as verdadeiras diferenças entre as sequências. Apesar de ocorrerem dois eventos de mutação na sequência derivada 1, apenas o último evento é observado, pois ocorreram no mesmo sítio. Os quadrados em vermelho evidenciam as diferenças em relação às sequências ancestrais.

genética indica uma relação evolutiva mais próxima, enquanto que um valor maior sugere uma derivação evolutiva proporcionalmente maior. Tipicamente, a informação da distância genética é incorporada à inferência filogenética na definição do tamanho dos ramos. No entanto, além desta informação é necessária uma escala de distância que especifique o número de mudanças que ocorreram ao longo do ramo.

O método mais simplista para avaliar a distância genética entre duas sequências é conhecido como distância p. Este método é baseado na contagem das diferenças dividida pelo número total de sítios do alinhamento. Se oito sítios são diferentes entre duas sequências homólogas com tamanho de 100pb, a distância p obtida será 0,08. Este resultado reflete a porcentagem de sítios diferentes em relação ao tamanho total da sequência, e geralmente é utilizado na especificação da escala de distância das filogenias (Figura 8-5).

A variação genética em um determinado sítio pode decorrer de diferentes processos e resultar em mais de uma substituição. As múltiplas substituições, ou *multiple hits*, ocorrem naturalmente e podem subestimar o verdadeiro número de mudanças no cálculo da distância p, já que "escondem" as diversas trocas de nucleotídeos ou aminoácidos. Na Figura 8-5b, por exemplo, apesar de ocorrerem duas substituições no mesmo sítio ao longo de um dos ramos, aparentemente a sequência derivada parece ter sofrido somente um evento evolutivo. Sendo assim, a relação entre as diferenças nas sequências e o tempo decorrido da divergência nem sempre é linear, especialmente devido à ocorrência das múltiplas substituições em um mesmo sítio.

Devido à ineficácia da distância p em efetivamente estimar a distância genética entre duas sequências, diferentes modelos probabilísticos foram desenvolvidos para descrever as mudanças entre os nucleotídeos e corrigir a distância observada. Tais modelos implicam no uso de diversas suposições simples a respeito das probabilidades de substituição de um nucleotídeo por outro, mas garantem uma aproximação da realidade quando sustentadas por uma taxa de mutação fidedigna.

Estas técnicas de correção são comumente conhecidas por modelos de substituição (ou matrizes de substituição), e garantem a conversão da distância observada em medidas de distâncias evolutivas próximas da realidade, permitindo reconstruir a história evolutiva dos organismos.

Diversos modelos de substituição foram propostos para explicar as trocas de nucleotídeos em sequências de DNA, reduzindo a complexidade do processo evolutivo a um padrão de mudança simples que consegue ser explicado através de poucos parâmetros. Todos estes modelos, no entanto, de alguma forma são inter-relacionados, diferindo principalmente no número de

parâmetros utilizados para explicar estas substituições. Devido à influência do modelo de substituição na inferência de filogenias, a escolha de um método particular deve ser justificada. A estratégia mais simples é utilizar os modelos que comportam o maior número de variáveis, embora a complexidade não esteja diretamente relacionada à melhor qualidade de análise das sequências. Com o aumento de parâmetros, o sistema se torna mais complexo, aumentando a probabilidade de erro e exigindo um maior processamento computacional. Assim, é necessário verificar os alinhamentos caso-a-caso para atribuir o melhor modelo de substituição na inferência filogenética.

A substituição de nucleotídeos ou aminoácidos em uma sequência é usualmente modelada sob a forma de um processo quase aleatório. Devido ao caráter dinâmico desta aleatoriedade, é necessário enquadrar as substituições, seguindo certos pressupostos. Assim, as substituições são descritas por um processo de Markov homogêneo, onde a probabilidade de substituição de um nucleotídeo X pelo Y não depende do estado prévio do nucleotídeo X.

As probabilidades de mudança de um nucleotídeo para outro (ou de um aminoácido para outro) são especificadas através de uma matriz 4x4 das taxas de substituição (ou 20x20 no caso dos aminoácidos) que especificam com qual taxa cada um dos nucleotídeos ou aminoácidos poderá mudar para outro. É necessário assumir também que os eventos de substituição sejam independentes ao longo dos sítios das sequências, e ainda, possuam um caráter reversível. Além disso, devem especificar a frequência estacionária dos nucleotídeos, ou frequência de equilíbrio, onde será atribuída a provável proporção de cada um dos caracteres na sequência.

Para sequências de nucleotídeos, o modelo de substituição mais simples foi proposto por Jukes e Cantor em 1969 (JC69). Segundo este modelo, as mudanças entre os nucleotídeos podem ocorrer com a mesma probabilidade, assumindo uma frequência estacionária igual para todos (cada nucleotídeo tem 25% de chance de ocorrer na sequência).

Com o advento da publicação das primeiras sequências de genoma mitocondrial, na década de 1980, se observou que as transições eram muito mais comuns que as transversões. Devido à uniformidade do método proposto por Jukes e Cantor, foi necessário criar um modelo que acomodasse essas diferenças.

Assim, o modelo proposto por Kimura (K80 ou K2P) cria as variáveis α e β para representar, respectivamente, as taxas de transição e de transversão. Apesar da inclusão de dois parâmetros, as frequências de equilíbrio se mantêm constantes em ¼ para cada nucleotídeo. Em 1981, Kimura adiciona um terceiro parâmetro (γ) ao modelo já proposto, passando a ser identificado como K3P. A atualização do modelo permitiu dividir as taxas de transversão em duas variáveis.

Alguns genomas apresentam uma grande quantidade de guaninas e citosinas em relação a timinas e adeninas. Se algumas bases são mais frequentes que outras, será esperado que algumas substituições ocorram com mais frequência que outras. O modelo criado por Felseinstein (F81) acomoda essas observações e permite que as proporções individuais de cada nucleotídeo (frequência estacionária) sejam diferentes de ¼. É importante ressaltar que este modelo considerará a mesma proporção de bases em todas as sequências envolvidas no alinhamento. Se diferentes sequências possuem diferente composição de bases, a pressuposição principal do modelo será violada.

O modelo HKY85, proposto por Hasegawa, Kishino e Yano, essencialmente mistura os modelos K2P e F81. Além de supor que a frequência das bases é variável, este modelo permite que transições e transversões ocorram com taxas diferentes.

Posteriormente, o modelo GTR (*generalised time-reversible*), o mais complexo dos modelos aqui apresentados, foi desenvolvido a partir do HKY85 com o intuito de acomodar diferentes taxas de substituição e diferentes frequências de bases. Este modelo requer seis parâmetros para taxa de substituição e quatro parâmetros para a frequência das bases, misturando todos os modelos aqui descritos.

Atualmente, além destes mais de 200 modelos de substituição podem ser aplicados a alinhamentos de nucleotídeos. Alguns programas, como Modeltest e Jmodeltest, são capazes de selecionar o modelo de substituição que melhor se ajusta a um dado alinhamento.

Uma importante extensão desses modelos de substituição incorpora a possibilidade de variação nas taxas evolutivas entre os sítios, permitindo ao modelo mais realismo. Assim, para cada sítio no DNA será atribuída uma probabilidade de evolução a uma taxa contida em um intervalo discreto de probabilidades. O método que garante a heterogeneidade de taxas evolutivas é modelado através de uma distribuição gama (Γ), que considera um número específico de taxas de

evolução para os sítios do DNA.

A aplicabilidade deste modelo nas inferências filogenéticas é facilitada pela simplicidade do método, já que apenas um único parâmetro (α) controla a forma da distribuição gama. Quando $\alpha < 1$, existe um grande número de taxas de evolução entre os sítios das sequências em análise, ou seja, quanto maior α, menor a heterogeneidade. Algumas vezes, uma proporção de sítios invariáveis (I), no qual uma determinada proporção de sítios é assumida como incapaz de sofrer substituição, pode também ser usada para modelar a heterogeneidade entre os sítios.

Ao contrário dos modelos de substituição de nucleotídeos, os modelos que explicam as trocas de aminoácidos são tradicionalmente empíricos. A partir da análise de alinhamentos de proteínas com identidade mínima de 85% Dayhoff, em 1970, desenvolveu uma série de matrizes de probabilidade que explicavam as mudanças de aminoácidos ao longo do tempo.

As matrizes PAM, como ficaram conhecidas, correspondem a modelos de evolução nos quais os aminoácidos são substituídos aleatoriamente e independentemente, de acordo com uma probabilidade predefinida que depende do próprio aminoácido.

Em 1992, um novo modelo de substituição de aminoácidos é criado por Henikoff e Henikoff. A análise de sequências de proteínas distantes evolutivamente, possibilitada pelo modelo de Henikoff-Henikoff, estabeleceu as bases para a criação das matrizes BLOSUM. As matrizes desta série foram identificadas por números (por exemplo, BLOSUM62) que se referem à porcentagem mínima de identidade dos blocos dos aminoácidos utilizados para construir o alinhamento. Matrizes similares, como GONNET e JTT, surgiram na mesma época.

Em 1996, foi proposto um modelo de substituição específico para proteínas codificadas pelo DNA mitocondrial, onde foi observado desvio de transições entre aminoácidos em relação às proteínas codificadas pelo material genético nuclear. Essa matriz, criada por Adachi e Hasegawa, foi chamada de mtREV.

Finalmente, em 2001, Whelan e Goldman propõem a matriz WAG, baseada em combinação e ampliação de vários modelos de substituição anteriores. Tal matriz é considerada superior às suas antecessoras para descrever filogenias de proteínas globulares.

5.5. Inferência filogenética

A reconstrução filogenética, ou seja, a reconstrução da história evolutiva de organismos, é um complexo processo que envolve uma série de etapas. O alinhamento, além de ser o primeiro passo, é um importante ponto para a inferência de filogenias (ver capítulo 3). Um alinhamento preciso, além de garantir maior confiabilidade nas análises posteriores, é requerido por todos os métodos de inferência filogenética para construção da árvore.

Depois que o alinhamento foi proposto, diversos métodos podem ser usados para estimar a filogenia das sequências estudadas. Podemos dividir estes métodos em dois principais grupos: métodos quantitativos e métodos qualitativos (Tabela 1-5). Estes grupos diferem na forma como os dados são tratados, refletindo diretamente como os dados do alinhamento serão inicialmente processados.

Os métodos quantitativos se baseiam na quantidade de diferenças entre as sequências do alinhamento para calcular uma árvore final. Já os métodos qualitativos constroem diversas filogenias que são classificadas seguindo uma determinada qualidade (critério). A filogenia que obtiver o maior valor associado à tal qualidade será a filogenia resultante.

Os métodos quantitativos compreendem os métodos de distância. Estes métodos convertem o alinhamento em matrizes de distância par-a-par para todas as sequências incluídas. Dentro destes algoritmos destacam-se dois métodos principais: UPGMA e aproximação dos vizinhos. Devido à grande eficiência computacional, estes métodos geralmente são utilizados para construção de uma filogenia inicial, que posteriormente é submetida a algum método do grupo qualitativo. Como principal ponto negativo, estes métodos apresentam apenas uma filogenia como resultado final (ver adiante).

Idealmente, todas as possíveis árvores para um dado alinhamento deveriam ser analisadas para garantir a escolha da melhor filogenia. Para isso, é necessário atribuir certos parâmetros que avaliem, dentre todas as ár-

Tabela 1-5: Comparação entre os tipos de métodos para inferência de filogenias.

Tipo	Método	Princípio	Programa
Métodos Quantitativos	UPGMA	Agrupa sequencialmente as OTUs com menor distância evolutiva entre si	Geneious MEGA
	Aproximação dos vizinhos	Busca a árvore com a menor soma total de ramos	MEGA Geneious HyPhy
Métodos Qualitativos	Máxima Parcimônia	Busca a filogenia com menor número de eventos evolutivos	PAUP MEGA Mesquite
	Máxima Verossimilhança	Busca a árvore com o valor de maior verossimilhança entre todas as filogenias construídas	PAUP PAML phyML MEGA
	Estatística Bayesiana	Amostra um número representativo de filogenias a partir do espaço amostral total de árvores e busca a mais provável	Mr. Bayes BEAST BAMBE

vores, aquela que explica as relações evolutivas de forma mais precisa.

Assim, os métodos qualitativos envolvem algoritmos que atribuem um critério de otimização para escolher a melhor filogenia. Nestes métodos, diversas filogenias são construídas e, seguindo um critério definido pelo algoritmo utilizado, uma filogenia será identificada como a que melhor explica a relação evolutiva entre os OTUs. O critério é utilizado para atribuir um valor a cada filogenia e ordená-las segundo este valor.

Estes métodos têm a vantagem de requerer uma função explícita para escolha das filogenias, sendo portanto independente da escolha do operador. No entanto, devido ao caráter de sua análise, são métodos mais refinados e intrinsecamente mais demorados computacionalmente. Três critérios de otimização são tradicionalmente empregados na inferência de filogenias: (a) Máxima Parcimônia, (b) Máxima Verossimilhança e (c) Inferência Bayesiana.

Por se tratarem de métodos que buscam uma única filogenia entre diversas árvores, os métodos qualitativos exigem algoritmos que vasculhem o maior número possível de filogenias em busca da melhor árvore. Dois grupos de algoritmos são destacados: os algoritmos exatos e os algoritmos heurísticos. Atualmente, devido ao tempo e à exigência computacional, os métodos heurísticos são preferidos aos exatos. No entanto, qualquer um deles pode ser aplicado aos métodos qualitativos de inferência filogenética. Como desvantagem dos métodos qualitativos, repetidos processos de procura em um mesmo conjunto de sequências podem levar a resultados diferentes, dependendo da árvore que é construída inicialmente pelo algoritmo.

Os métodos exatos buscam todas as filogenias possíveis para um grupo de sequências. O funcionamento destes métodos geralmente envolve a seleção aleatória inicial de três OTUs para a construção de uma árvore filogenética não enraizada. Por tentativa, um a um, novas OTUs, também tomadas aleatoriamente do alinhamento, são inseridas em diferentes posições na árvore. Esse procedimento é repetido até todos os táxons serem inseridos, garantindo que todas as filogenias possíveis para o alinhamento dado sejam geradas.

A partir da aplicação de um critério de otimização (dado pelo método qualitativo) para classificar as filogenias e ordená-las segundo este valor, é possível organizar um espaço virtual que contém todas as filogenias possíveis para o alinhamento empregado. É importante lembrar que, tomando poucas sequências, milhões de árvores podem ser geradas. Este conjunto total de filogenias é comumente chamado de espaço amostral. Como exemplo, podemos organizar o espaço amostral de filogenias originadas a partir de um alinhamento de dez sequências em um gráfico bidimensi-

onal baseado no valor atribuído pelo critério de otimização a cada árvore (Figura 9-5). Nestas condições, será possível observar que algumas árvores possuem valores maiores que outras, formando picos que agrupam as melhores filogenias. Da mesma forma, entre diferentes picos existem vales representados por árvores com valores menores e, portanto, menos consistentes.

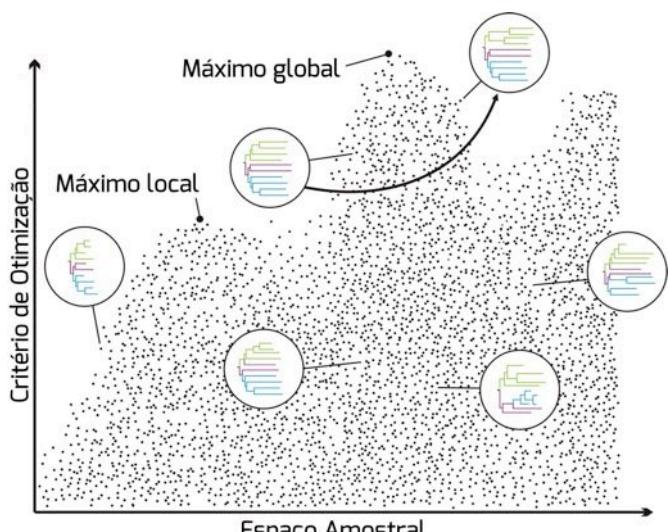



Figura 9-5: Descrição de parte do espaço amostral das possíveis filogenias para um determinado sistema, ordenadas segundo um valor atribuído pelo critério de otimização. Cada ponto no gráfico representa uma topologia diferente inferida a partir de um conjunto de dez sequências homólogas. O espaço amostral, neste caso, é definido por 2.027.025 filogenias e apresenta, segundo o critério de otimização, dois máximos locais e um máximo global, que contém as melhores filogenias. Em destaque, algumas filogenias exemplificando as possibilidades de arranjo dos ramos. A seta indica a mudança de topologia da filogenia e o consequente aumento de seu valor dado pelo critério de otimização.

Os métodos de busca exaustiva construirão um espaço amostral de árvores através de métodos específicos de modificação das filogenias. Por acumularem um grande número de resultados, estes métodos exigem um tempo computacional muito elevado, por vezes tornando-se proibitivos.

Os algoritmos de busca heurística procuram pela melhor filogenia em um subconjunto de todas as filogenias possíveis. Apesar de serem muito mais rápidos computacionalmente, estes métodos não garantem que a filogenia correta seja encontrada, pois apenas algumas árvores do espaço amostral total serão consideradas. Ainda assim, estes métodos tem mostrado grande eficiência.

Atualmente, os principais métodos qualitativos de inferência filogenética incorporam algoritmos de busca heurística para amostrar as filogenias do espaço amostral virtual. Usualmente, estes algoritmos de busca são executados em dois passos. Primeiramente, diferentes árvores são construídas e, após encontrar a melhor árvore guiada por um critério de otimização, aplica-se um algoritmo para modificar aleatoriamente o arranjo dos ramos. Este método permite testar se outros arranjos são ou não mais consistentes.

Devido ao grande número de métodos para inferência filogenética, a decisão quanto ao uso de cada um é de grande importância para a interpretação do resultado final: a filogenia. Ao escolher um método, é fundamental verificar o poder (tamanho e quantidade de sequências necessária para resolver a filogenia), a eficiência (habilidade de estimar a filogenia correta com um número limitado de dados), a consistência (habilidade de estimar a filogenia correta com um número de dados ilimitado) e a robustez (habilidade de estimar a filogenia correta quando certos pressupostos da análise são violados).

Até o momento, não existe um método que apresente todas estas características simultaneamente e garanta a reconstrução filogenética correta. É importante, sobretudo, conhecer a biologia do organismo (ou dos organismos) em questão para que a escolha do método tenha, além de tudo, uma justificativa biológica.

5.6. Abordagens quantitativas

UPGMA

O método baseado em distâncias UPGMA (*unweighted pair-group method using arithmetic averages*, ou método de agrupamento par a par usando médias aritméticas não ponderadas) foi proposto por Sneath e Sokal, em 1973, e é o método mais simples para reconstrução filogenética. O UPGMA

parte do pressuposto de que todas as linhagens evoluem a uma taxa constante (hipótese do relógio molecular).

No UPGMA, uma medida de distância evolutiva é computada para todos os pares de sequências utilizando um modelo evolutivo. Após, estas distâncias são organizadas na forma de uma matriz, conforme ilustrado abaixo:

Sequências	1	2	3	4
2	$d_{1,2}$			
3	$d_{1,3}$	$d_{2,3}$		
4	$d_{1,4}$	$d_{2,4}$	$d_{3,4}$	
5	$d_{1,5}$	$d_{2,5}$	$d_{3,5}$	$d_{4,5}$

O agrupamento das sequências é iniciado pelo par com menor distância. Supondo que $d_{1,2}$ seja a menor distância no exemplo acima, as sequências *1* e *2* são agrupadas com um ponto de ramificação na metade dessa distância ($d_{1,2/2}$). As sequências *1* e *2* são então combinadas em uma entidade composta, agora denominada y, e a distância entre esta entidade y e as outras sequências é computada (observe abaixo).

Sequências	$y_{(1,2)}$	3	4
3	$d_{y,3}$		
4	$d_{y,4}$	$d_{3,4}$	
5	$d_{y,5}$	$d_{3,5}$	$d_{4,5}$

Supondo que $d_{y,3}$ seja a menor distância, y e *3* são combinados em uma nova entidade composta, digamos, z. Seu ponto de ramificação é calculado levando em conta a distância de cada membro de y (*1* e *2*) em relação a *3* e dividindo por *2*, ou seja, $(d_{1,3}+d_{2,3})/2$. O mesmo procedimento se repete, calculando a menor distância entre z e outra sequência (suponhamos que seja a sequência *4*). Calculam-se a distância de cada membro de z até *4*, divide-se o somatório das distâncias por dois e cria-se uma nova sequência composta. O mesmo procedimento é repetido até que existam apenas duas sequências a serem agrupadas (comumente, uma sequência simples e uma entidade composta).

Ao empregar sequências de DNA ou proteína proximamente relacionadas, o UPGMA pode construir duas ou mais "árvores empatadas" (*tie trees*). Essas árvores surgem quando dois ou mais valores de distância na matriz se mostram idênticos. É possível representar todas as árvores empatadas, mas essa abordagem é pouco útil, uma vez que tais árvores são muito semelhantes e surgem por erros de estimativa das distâncias. Para tais casos, sugere-se apresentar uma única árvore, geralmente a árvore consenso do *bootstrap* (ver seção 5.8).

Por se basear na hipótese do relógio molecular, o UPGMA pode levar à obtenção de topologias falsas quando tal hipótese não for satisfeita pelos dados. Sabe-se que o método é muito sensível a variações nas taxas evolutivas entre linhagens, fato este que levou a proposição de métodos onde as variações são ajustadas para a obtenção de sequências que satisfaçam o relógio molecular. Apesar disso, devido ao surgimento de métodos mais robustos e mais eficientes em lidar com dados não uniformes, o UPGMA encontra-se praticamente abandonado como alternativa para reconstrução filogenética.

Aproximação dos Vizinhos

O método de aproximação dos vizinhos (*neighbor joining* ou NJ) foi proposto por Saitou e Nei em 1987. Este método se baseia em um aceleramento dos algoritmos de evolução mínima que existiam até então. Em sua versão original, estes algoritmos buscavam a árvore com menor soma total de ramos, de maneira que todas as árvores possíveis precisavam ser construídas para que se verificasse qual delas apresentava a menor soma. O algoritmo de NJ facilitou esse processo, tendo o princípio de evolução mínima implícito no processo e produzindo apenas uma árvore final.

Para construir a filogenia, o NJ começa por uma árvore totalmente não resolvida (topologia em estrela) (Figura 10-5). Tendo como base uma matriz de distâncias (semelhante à matriz inicial construída pelo método de UPGMA) entre todos os pares de sequências, construída a partir da aplicação de um modelo de substituição (conforme descrito na seção 5.4), o par que apresentar a menor distância é identificado, unido por um nó (que representará o ancestral comum deste par de sequências) e incorporado na árvore (na Figura 10-5, f e g são unidos pelo nó u). As distâncias de cada sequência do par são recalculadas em relação ao novo nó u, assim como as distâncias de todas as outras sequências são recalculadas em relação ao novo nó u. O algoritmo reinicia, substituindo o par de vizinhos unidos pelo novo nó e usando as distâncias calculadas no passo anterior.

Quando duas somatórias de ramos são iguais, a decisão sobre quais ramos unir depende do programa empregado. Alguns optam pela primeira sequência apresentada no arquivo de dados, enquanto outros escolhem aleatoriamente qual dos pares deve ser unido primeiro. Árvores empatadas (*tie trees*) são raras com o uso de NJ, e recomenda-se o emprego da árvore consenso do *bootstrap* (ver seção 5.8) para evitá-las. Uma variação do algoritmo NJ, o BIONJ tem se mostrado ligeiramente melhor que o NJ em casos pontuais; no entanto, conserva o mesmo princípio do algoritmo.

5.7. Abordagens qualitativas

Parcimônia

O princípio de parcimônia foi proposto por Guilherme de Occam (ou *William of Ockham*) no século XVII. Occam defendia que a natureza é por si só econômica e opta por caminhos mais simples. O pensamento se espalhou por diversas áreas do conhecimento e, atualmente, seu princípio é conhecido como Navalha de Occam.

Historicamente, a parcimônia teve um papel muito importante no estabelecimento da disciplina de filogenética molecular. Desde 1970, foi o critério de otimização mais utilizado para inferência de filogenias.

Contudo, atualmente a máxima parcimônia foi substituída por outros métodos, como máxima verossimilhança e inferência Bayesiana devido, principalmente, às simplificações nos processos evolutivos assumidas pelo método e, sobretudo, nas limitações de seu uso. Apesar disso, a máxima parcimônia ainda está integrada ao campo da inferência filogenética por ser um método rápido e, em alguns casos, muito efetivo.

A aplicação do princípio de máxima parcimônia nas reconstruções filogenéticas é conceitualmente simples: dentro de um conjunto de filogenias, aquela filogenia que apresentar o menor número de eventos evolutivos (substituições) deve ser a mais provável para explicar os dados do alinhamento.

Figura 10-5: Começando com uma árvore em estrela (a), a matriz de distâncias é calculada para identificar o par de nós a ser unido (nesse caso, f e g). Estes são unidos ao novo nó u (b). A porção em vermelho é fixada e não será mais alterada. As distâncias do nó u até os nós a-e são calculadas e usadas para unir o próximo vizinho. No caso, u e e são unidos ao recém criado nó v (c). Mais duas etapas de cálculo levam à árvore em (d) e então à árvore em (e), que está totalmente resolvida, encerrando o algoritmo.

Metodologicamente, o critério de parcimônia deve determinar a quantidade total de mudanças na filogenia, descrevendo o tamanho dos ramos. Adicionalmente, a parcimônia guia a busca, entre todas as árvores possíveis, daquela filogenia que minimiza os passos evolutivos de forma máxima sendo, portanto, a filogenia de máxima parcimônia.

Assim que uma determinada filogenia é proposta, o método calculará as probabilidades de mudanças dos nucleotídeos desde os ramos terminais até os ramos mais ancestrais da árvore. Por se tratar de um método qualitativo, a parcimônia considera cada sítio do alinhamento individualmente e calcula as probabilidades de ocorrência dos quatro nucleotídeos nos táxons ancestrais.

Devido ao caráter probabilístico do método, é necessário que certas pressuposições sejam estabelecidas para especificar o custo de substituição dos nucleotídeos. A forma mais simples do método (Parcimônia de Wagner) assume que as substituições de nucleotídeos tem custo 1, enquanto que a não alteração não é penalizada (Figura 11-5a). No entanto, esquemas um pouco mais complexos que levam em consideração as questões biológicas envolvidas no processo evolutivo foram propostas. Um esquema comum de matriz com custo desigual, proposto para especificar as transições e as transversões, leva em consideração a diferença na probabilidade de mudança entre purinas e pirimidinas (Figura 11-5b). Comumente, a matriz é especificada sem que constem os respectivos nucleotídeos, no entanto, por convenção são atribuídos nas linhas e colunas em ordem alfabética (A, C, G e T).

Para o método de parcimônia, apenas sítios variáveis são considerados informativos. Estes sítios devem apresentar dois caracteres diferentes presentes em, no mínimo, dois indivíduos (Figura 12-5b). Aqueles sítios que não apresentam variação ou apresentam autapomorfias (caracter diferente presente em apenas um indivíduo) serão descartados automaticamente das análises.

Devido ao tamanho dos alinhamentos e ao número de OTUs incluídas para a inferência de filogenias, foi necessário que algoritmos fossem desenvolvidos para acelerar os cálculos na busca pela árvore de máxima parcimônia. Algoritmos de programação dinâmica são capazes de lidar com a atribuição de custos e realizar os devidos cálculos para escolha da filogenia com o menor custo. Diversos algoritmos foram desenvolvidos, embora a parcimônia de Sankoff, desenvolvida em 1975, tenha se tornado uma das mais populares.

a.

Matriz de custo igual =	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

b.

Matriz de custo desigual =	A	C	G	T
A	0	4	1	4
C	4	0	4	1
G	1	4	0	4
T	4	1	4	0

Figura 11-5: Matrizes de custo aplicadas ao método de máxima parcimônia para penalizar as substituições de um nucleotídeo por outro. (a) Matriz de custos iguais para todas as mudanças entre nucleotídeos. (b) Matriz de custo desigual, considerando a maior probabilidade de ocorrência de transições em relação às transversões ao longo do processo evolutivo.

Após a atribuição de uma matriz de custo e a proposição de uma filogenia, o algoritmo utilizará cada um dos sítios informativos do alinhamento independentemente para cálculo dos custos (Figura 11-5).

Considere a matriz desigual da Figura 11-5b e a filogenia inicialmente proposta na Figura 12-5a. O esquema demonstra que para cada sítio informativo será construída uma filogenia com a mesma topologia da árvore proposta em 12-5a (ver adiante).

Tomando, por exemplo, o sítio 28, identificamos a presença de três ancestrais não amostrados que, no entanto, para o cálculo dos custos, terão que ter seus caracteres inferidos. Segundo o algoritmo de Sankoff, os cálculos devem iniciar tomando os clados mais derivados (isto é, mais recentes). Em 12-

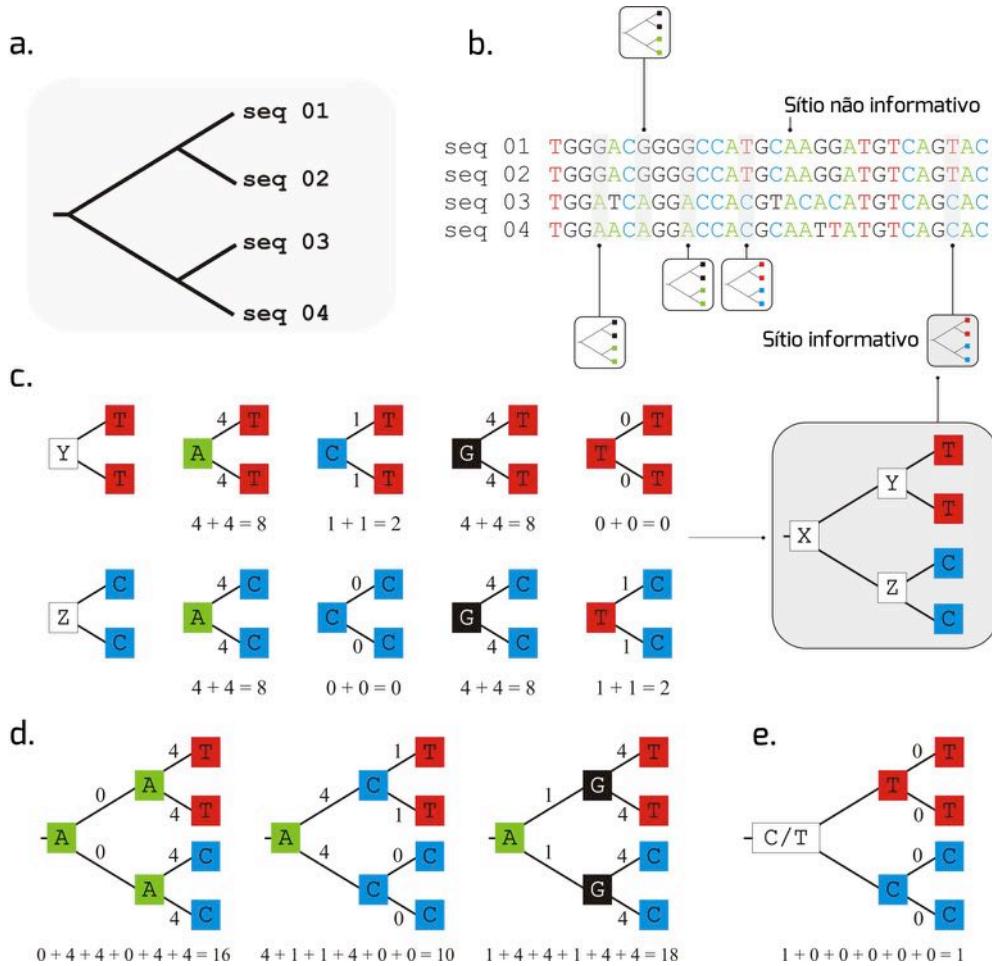



Figura 12-5: Determinação dos custos de substituição pelo método de parcimônia para um sítio do alinhamento de nucleotídeos. (a) Topologia da filogenia proposta para quatro táxons (ver adiante). (b) Alinhamento de nucleotídeos de quatro sequências homólogas. Destacados em cinza estão os sítios informativos para o método de parcimônia. Os demais sítios são considerados não informativos e serão descartados durante os cálculos. (c) Cálculo dos custos para os dois clados presentes na filogenia proposta em "a". O método supõe que a posição "Y" possa ser ocupada por qualquer um dos quatro nucleotídeos. (d) Exemplo do procedimento adotado pelo método, supondo que a posição "X" na filogenia foi ocupada pelo nucleotídeo A. É necessário considerar todas as possibilidades de caracteres nos sítios ancestrais e calcular os respectivos custos. (e) Arranjo de menor custo para a posição 28 do alinhamento de nucleotídeos.

5c, a posição "Y" da filogenia necessariamente foi ocupada por um dos quatro nucleotídeos. Em cada uma das proposições (A, C, G ou T), o custo associado à substituição é consultado na matriz. No primeiro caso, a hipótese para ocupação da posição "Y" é A. O custo da substituição em cada um dos ramos deve ser verificado e somado. Por exemplo, a substituição de A por T possui custo 4. Como a mesma substituição ocorreu em dois ramos diferentes, somamos o custo total, que totaliza 8. O mesmo procedimento será repetido considerando os outros três nucleotídeos na posição "Y".

Após o cálculo dos custos para as posições "Y" e "Z", é necessário verificar os custos de substituição de "X" para "Y" e "X" para "Z". A Figura 12-5d apresenta a primeira hipótese para ocupação da posição "X": o nucleotídeo A. Aqui, o algoritmo somará os custos de substituição de todos os ramos, novamente considerando cada um dos quatro

nucleotídeos na posição “X”, mas também considerando a variação nas posições “Y” e “Z”. A Figura 12-5e identifica a filogenia com o menor custo para o sítio 28. Note que o caractere mais ancestral pode ser tanto o nucleotídeo T quanto C. Os mesmos cálculos serão realizados para todos os sítios do alinhamento, tomando a topologia dada em 12-5a e, ao final, os menores custos para cada sítio serão somados para encontrar o tamanho dos ramos da árvore. A árvore que possuir os ramos mais parcimoniosos será tomada como a árvore de máxima parcimônia.

Computacionalmente, o cálculo dos tamanhos de ramos mais parcimoniosos não é um problema. O desafio da maioria dos métodos de reconstrução filogenética está na inferência da topologia. Assim como no método de máxima verossimilhança, discutido a seguir, o método de máxima parcimônia contará com algoritmos heurísticos para arranjo das topologias. A filogenia é então proposta pelo algoritmo, e o critério de parcimônia avalia a árvore. A partir de perturbações realizadas nesta topologia, uma nova topologia é proposta e novamente o critério qualifica a filogenia.

Apesar de velozes, os métodos de parcimônia falham ao estimar a relação evolutiva entre um grande número de táxons, especialmente se diferentes linhagens possuem taxas evolutivas variáveis ou taxas evolutivas muito rápidas. Nestes casos, é comum que o método agrupe incorretamente os táxons com maiores taxas de evolução, levando à inferência da filogenia errada (atração de ramos longos).

Ainda, por não ter um modelo de substituição especificado, o método de parcimônia é incapaz de considerar mutações reversas ou múltiplas substituições. Métodos que geram diferentes hipóteses a partir do alinhamento, considerando as observações biológicas na seleção do modo de substituição dos nucleotídeos e, assim, lidam com eventos aleatórios de probabilidade, substituíram o uso da máxima parcimônia e, atualmente, são os principais métodos utilizados para a inferência de filogenias.

Máxima Verossimilhança

Idealmente, os métodos de inferência filogenética devem resgatar o máximo de informações contidas em um dado conjunto de sequências homólogas, buscando desvendar a verdadeira história evolutiva dos organismos.

Quando um grande número de mudanças evolutivas em diferentes linhagens é demasiadamente desigual, o método de máxima parcimônia tende a inferir filogenias inconsistentes, proporcionalmente convergindo à árvore errada quanto maior o número de sequências no alinhamento. Assim, abre-se espaço para uma técnica de inferência filogenética mais robusta, que alie as informações do alinhamento a um modelo estatístico capaz de lidar com a probabilidade de mudança de um nucleotídeo para outro de maneira mais completa.

Dentro do campo da filogenética computacional, o método de máxima verossimilhança primeiramente ocupou este espaço e, desde então, tem sido amplamente utilizado devido à qualidade da abordagem estatística empregada.

A implementação de uma concepção estatística para a máxima verossimilhança, originalmente desenvolvida para estimar parâmetros desconhecidos em modelos probabilísticos, se deu entre 1912 e 1922 através dos trabalhos de A. R. Fisher.

Apesar de utilizado para dados moleculares na década de 1970, o método de máxima verossimilhança só se tornou popular na área da filogenética a partir de 1981, com o desenvolvimento de um algoritmo para estimar filogenias baseadas no alinhamento de nucleotídeos. Atualmente, diversos programas implementam este método para realizar a inferência filogenética, incluindo PAUP, MEGA, PHYLIP, fastDNAml, IQPNNI e METAPIGA, dentre outros (Tabela 1-5).

O objetivo principal do método da máxima verossimilhança é inferir a história evolutiva mais consistente com relação aos dados fornecidos pelo conjunto de sequências. Neste

modelo, a hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) é avaliada pela capacidade de predizer os dados observados (alinhamento de sequências homólogas). Sendo assim, a verossimilhança de uma árvore é proporcional à probabilidade de explicar os dados do alinhamento. Aquela árvore que com maior probabilidade, entre as outras árvores possíveis, produz o conjunto de sequências do alinhamento, é a árvore que reflete a história evolutiva mais próxima da realidade, mais verossímil e, por isso, de máxima verossimilhança.

É importante ressaltar que diferentes filogenias podem explicar um determinado conjunto de sequências, algumas com maior probabilidade e, outras, com menor probabilidade. No entanto, a soma das verossimilhanças de todas as árvores possíveis para um determinado conjunto de sequências nunca resultará em 1, pois não estamos lidando com as probabilidades de que estas filogenias estejam corretas, mas avaliando a probabilidade de explicarem o alinhamento que foi fornecido.

Se, por exemplo, aplicássemos o método de máxima verossimilhança para inferir a árvore filogenética de um grupo de sequências homólogas que incluem porções recombinantes, encontraríamos uma árvore filogenética com um determinado valor de verossimilhança. A utilização do método, por si só, garantiria como resultado a inferência de uma filogenia. No entanto, sabemos que esta árvore, apesar de ser a mais plausível para explicar o alinhamento dado, não tem qualquer relação com a realidade evolutiva do organismo, já que eventos de recombinação aconteceram no decorrer do tempo e impedem a explicação sob a forma dicotômica de uma filogenia.

A aplicação do método de máxima verossimilhança exige a construção de uma filogenia inicial, geralmente obtida por métodos quantitativos. Como exemplo, considere a árvore filogenética proposta inicialmente e o respectivo alinhamento de nucleotídeos da Figura 13-5. Para calcularmos a verossimilhança desta filogenia será necessário utilizar um modelo evolutivo, que será importante para atribuir valores e parâmetros às substituições e ajudará no cálculo da probabilidade de que uma sequência X mude para uma sequência Y ao longo de um segmento da árvore.

Dado um determinado modelo evolutivo (JC69, K2P, F81, HKY ou GTR, por exemplo), e assumindo que cada sítio do alinhamento evolui de maneira independente dos demais, podemos calcular o valor de verossimilhança para cada um destes sítios e, posteriormente, multiplicar os valores de cada sítio para encontrar a verossimilhança da árvore dada (Figura 13-5 e a Figura 14-5). Sítios que apresentam deleções serão eliminados da análise.

Como os nós internos destas árvores, geradas a partir de cada sítio do alinhamento, são a representação de OTUs não amostrados (isto é, ancestrais) e, por conseguinte, não se conhecem suas sequências de nucleotídeos, será necessário considerar a ocorrência de todos os nucleotídeos (A, T, C e G) nestas posições da árvore (Figura 13-5c).

Por certo, alguns cenários são mais prováveis que outros; no entanto, todos devem ser considerados durante os cálculos de verossimilhança, pois apresentam alguma probabilidade de terem gerado as sequências dadas no alinhamento. Adicionalmente, além de calcular a probabilidade de todas as mudanças possíveis para cada um dos sítios do alinhamento (Figura 13-5c), a expressão matemática da verossimilhança ainda incluirá o tamanho dos ramos, dentre outros elementos do modelo de substituição, como um fator determinante para o cálculo (Figura 13-5d).

A probabilidade de ocorrência de cada um dos quatro nucleotídeos no nó mais interno da árvore será igual à respectiva frequência estacionária dada pelo modelo de substituição, já que este parâmetro especifica a proporção esperada de cada um dos quatro nucleotídeos. No modelo de Jukes e Cantor, por exemplo, assume-se que os quatro nucleotídeos ocorrem em proporções iguais de 25%.

Conforme o exemplo da Figura 13-5d, a equação utilizada para calcular a verossimilhança da filogenia

proposta no sítio 28, inicialmente, leva em consideração a frequência estacionária do nucleotídeo G, já que este é o nucleotídeo que está sendo considerado como presente no nó mais ancestral da árvore. A probabilidade de este G ser substituído por um A (P_{GA}), ou permanecer G (P_{GG}) será dada pelo modelo de substituição escolhido. Da mesma forma, serão os casos P_{GT}, P_{AC} (repetido duas vezes cada pelo fato de existirem dois ramos terminais com o mesmo nucleotídeo).

O tamanho dos ramos entre dois nós será multiplicado pelas probabilidades de substituição dos nucleotídeos, levando em conta variações em parâmetros do modelo de substituição. Apesar da dificuldade de cálculo computacional, os algoritmos aplicados à inferência filogenética (baseados no princípio de Pulley) automaticamente estimarão o tamanho de cada ramo de modo que este maximize o valor da verossimilhança da árvore filogenética em construção. Nestes casos, o algoritmo atribui diversos valores de distância para um ramo e, a cada valor, verifica a verossimilhança da árvore, buscando aqueles valores que resultam na filogenia com a maior verossimilhança.

A probabilidade de observar os dados em um sítio particular é a soma das probabilidades de todos os possíveis nucleotídeos que poderiam ser observados nos nós internos da árvore (Figura 13-5c). O número de

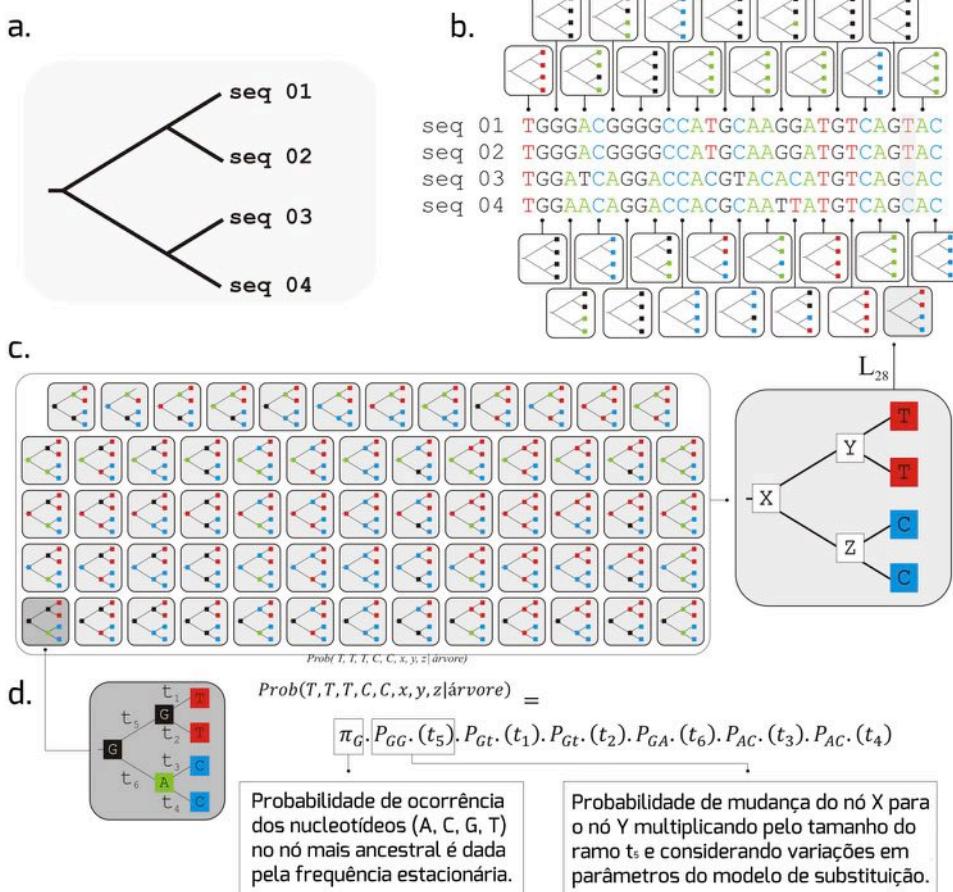



Figura 13-5: Esquema do cálculo da verossimilhança para uma filogenia e seu respectivo alinhamento de nucleotídeos. (a) Árvore filogenética proposta inicialmente para o alinhamento em “b”. (b) Para cada posição do alinhamento é destacada a organização dos quatro sítios do alinhamento na árvore proposta em “a”. Como exemplo, apenas o sítio do alinhamento destacado em cinza será considerado para o cálculo da verossimilhança. Os quadrados pretos, azuis, verdes e vermelhos nos ramos terminais das filogenias representam, respectivamente, os nucleotídeos guanina, citosina, adenina e timina. (c) Probabilidade de cada uma das 64 possíveis combinações de nucleotídeos nos nós internos da árvore, já que estes representam os sítios de táxons ancestrais não amostrados (P_{XY}, P_{YT}, P_{XZ}, P_{ZC}). (d) O esquema para o cálculo da máxima verossimilhança leva em conta a multiplicação do tamanho dos ramos (t_1, t_2, t_3, t_4, t_5 e t_6) pelas respectivas probabilidades de transição (P_{GG}, P_{GT}, P_{GA} e P_{AC}), além da frequência estacionária dos quatro nucleotídeos no nó mais ancestral (π_X).

nós internos rapidamente se torna muito grande com o aumento do número de OTUs. Felizmente, através de um algoritmo criado por Felseinstein (algoritmo de "poda"), que se aproveita da própria topologia da filogenia, esses cálculos podem ser realizados de uma maneira computacionalmente eficiente.

Neste processo, propõe-se que os cálculos da verossimilhança de uma determinada árvore sejam feitos a partir de sub-árvores dos ramos terminais em direção aos nós internos, semelhante ao algoritmo usado para o cálculo da parcimônia. No entanto, quando aplicado este método à inferência por máxima verossimilhança é necessário garantir que os modelos de substituição, não presentes no método de máxima parcimônia, sejam reversíveis, ou seja, que a probabilidade de mudança de A para T (P_{AT}) seja a mesma que T para A (P_{TA}). A introdução deste método permitiu que as análises de verossimilhança pudessem ser aplicadas a grandes conjuntos de sequências, de forma mais rápida e efetiva.

Ao final, multiplicamos os valores de verossimilhança de todos os sítios e encontramos o valor de verossimilhança da árvore (Figura 14-5):

A expressão matemática acima indica que a verossimilhança (L) é igual à multiplicação (Π) das probabilidades de cada sítio i (D^i, calculado conforme Figura 13-5), dada a árvore filogenética (topologia, modelo evolutivo e tamanho dos ramos). Aquela árvore que tiver o maior valor de verossimilhança entre todas as árvores possíveis para um determinado alinhamento de sequências será a árvore que melhor explica o alinhamento e, por isso, a árvore de máxima verossimilhança. Por fim, é importante ressaltar que, apesar de estarmos avaliando nucleotídeos neste exemplo, o mesmo raciocínio poderia ser aplicado para a inferência filogenética para um alinhamento de aminoácidos.

Até o momento vimos, em linhas gerais, como realizar o cálculo de verossimilhança para uma dada filogenia (Figura 13-5). No entanto, outra função importante dos métodos computacionais de inferência filogenética é apontar a topologia e encontrar a árvore de máxima verossimilhança entre todas as árvores possíveis para o conjunto de dados. Infelizmente, não existem algoritmos que garantam a localização da árvore real devido ao grande espaço amostral de árvores possíveis (Figura 9-5).

Após uma árvore ser construída, é necessário calcular sua verossimilhança e comparar este valor com todas as árvores já construídas. Como é impossível testar a verossimilhança para todas as filogenias possíveis, os algoritmos de máxima verossimilhança incluirão buscas heurísticas para solucionar este problema (estes métodos construirão diferentes filogenias a partir do mesmo conjunto de dados do alinhamento).

Figura 14-5: Cálculo da máxima verossimilhança de uma dada filogenia, considerando seu respectivo alinhamento de nucleotídeos contendo quatro táxons e 30 bases (Figura 13-5b). Para cada sítio (L_{01}, L_{02}, ..., L_{30}) será calculado um valor de probabilidade que envolve a consideração de todos os quatro nucleotídeos em cada um dos ramos ancestrais da filogenia. Posteriormente, os valores de verossimilhança de cada sítio serão multiplicados para encontrar a verossimilhança total da filogenia.

Na problemática das filogenias, diferentes programas têm proposto as mais diversas alternativas para avaliar o maior número de árvores do espaço amostral total e encontrar aquela com o maior valor de verossimilhança. No entanto, como regra geral, a maioria dos programas de máxima verossimilhança segue alguns passos comuns:

i) Uma filogenia preliminar com determinada topologia é construída (geralmente são utilizadas árvores construídas pelo método de aproxima-

ção de vizinhos);

ii) Os parâmetros para esta árvore são modificados buscando maximizar a verossimilhança (em alguns casos, a filogenia vai sendo construída pela adição de novos táxons aleatoriamente). Para a modificação da filogenia, os algoritmos podem implementar técnicas de rearranjos de ramos, conforme descrito em 5.4;

iii) O valor de máxima verossimilhança para esta árvore é armazenado;

iv) Outras topologias são construídas e seus parâmetros também são avaliados;

v) Finalmente, a filogenia que possuir o valor de máxima verossimilhança será a melhor estimativa evolutiva para o dado conjunto de sequências.

Embora estes processos simplifiquem os verdadeiros fenômenos biológicos que governam a evolução de uma sequência, apresentando assim dificuldades em identificar a árvore com o maior valor de verossimilhança, eles são normalmente robustos o bastante para estimar as relações evolutivas entre táxons.

Como estes métodos implicam em encontrar a árvore com o valor máximo de verossimilhança entre todas as árvores amostradas, o resultado final sempre fornecerá apenas uma filogenia, ao contrário dos métodos Bayesianos que serão vistos a seguir. Cabe ressaltar que, devido ao uso de diferentes algoritmos, na prática, um mesmo conjunto de sequências submetido a diferentes programas para inferência filogenética por máxima verossimilhança dificilmente resultará na mesma árvore. Por isso, é necessário ser cauteloso ao interpretar árvores geradas pelo método de máxima verossimilhança.

Análises Bayesianas

A estatística Bayesiana nasceu com a publicação de um ensaio matemático do reverendo Thomas Bayes, em 1793. Nesta publicação, o reverendo apresenta o desenvolvimento de um método formal para incorporar evidências prévias no cálculo da probabilidade de acontecimento de determinados eventos.

Inicialmente, este método foi aplicado apenas no campo da matemática e, só a partir de 1973, passa a ser incorporado no pensamento biológico e na inferência filogenética. Com o advento de diversos programas de acesso livre para realizar a inferência de filogenias por estatística Bayesiana, o método se difundiu e, atualmente, tornou-se um campo de estudo específico dentro da filogenética computacional.

A inferência Bayesiana engloba o método de máxima verossimilhança (Tabela 2-5) mas, adicionalmente, inclui o uso de informações dadas *a priori*. Estas informações refletem características a respeito da filogenia, do alinhamento ou dos táxons, que o pesquisador sabe de antemão.

Entre os principais parâmetros que podem ser conhecidos antes da reconstrução filogenética pode-se destacar a taxa evolutiva, tipo de relógio molecular, parâmetros do modelo de substituição, datas de coleta das amostras, datas para calibração da filogenia (achados fósseis, datação por carbono-14, aproximações arqueológicas, etc.), distribuição geográfica, organização monofilética de um grupo de indivíduos ou, até mesmo, parâmetros de dinâmica populacional.

Os valores atribuídos *a priori* são incorporados à estatística Bayesiana na forma de probabilidades e comporão o termo chamado de probabilidade anterior (*prior probability*). Se sabemos de antemão que um determinado grupo de organismos é ancestral em relação a outro, podemos atribuir uma maior probabilidade àquelas filogenias que relacionam estes organismos da maneira como sabemos *a priori*.

Qualquer informação útil, que é fornecida pelo pesquisador antes da própria reconstrução da filogenia, poderá ser convertida em uma probabilidade anterior para ser inserida nas análises de inferência Bayesiana. No entanto, as informações cedidas *a priori* devem

Tabela 2-5: Comparação entre os métodos de máxima verossimilhança e inferência Bayesiana.

Método	Vantagens	Desvantagens
Máxima Verossimilhança	Captura totalmente a informação dos sítios do alinhamento para construção das filogenias	Comparativamente ao método Bayesiano, o algoritmo para reconstrução por máxima verossimilhança é mais lento
Estatística Bayesiana	Tem grande ligação com a máxima verossimilhança, sendo, no entanto, geralmente mais rápida. Modelos populacionais podem ser incluídos para inferência das filogenias	Os parâmetros para as probabilidades anteriores devem ser especificados e pode ser difícil especificar quando as análises são satisfatórias

ser distribuições de números prováveis (mínimo e máximo), e não números exatos. Quando estes valores não são conhecidos ou quando, por exemplo, não se quer atribuir maior probabilidade a uma determinada topologia, o parâmetro terá uma distribuição uniforme de probabilidades.

Na maioria dos aplicativos que lidam com inferência Bayesiana existem distribuições uniformes associadas às probabilidades anteriores que assumem que todos os valores possíveis são dados pela mesma probabilidade.

Além das probabilidades anteriores, a inferência Bayesiana é baseada nas probabilidades posteriores de um parâmetro como, por exemplo, a topologia. Através da probabilidade posterior é possível verificar a probabilidade de cada uma das hipóteses (árvores filogenéticas). Sendo assim, ao final das análises, é possível estabelecer uma estimativa da probabilidade dos eventos retratados por uma determinada filogenia, ou seja, a probabilidade de cada filogenia. As probabilidades posteriores são calculadas utilizando a fórmula de Bayes:

$$\mathrm{L}(H \mid D) = \frac{\mathrm{L}(H)\,\mathrm{L}(D \mid H)}{\mathrm{L}(D)}$$

O termo $\mathrm{L}(H \mid D)$ é chamado de distribuição de probabilidades posteriores, e é dado pela probabilidade da hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) a partir dos dados disponíveis (alinhamento de sequências). O termo $\mathrm{L}(D \mid H)$ descreve o cálculo de máxima verossimilhança, enquanto o multiplicador $\mathrm{L}(H)$ é a probabilidade anterior. Para o termo que envolve a função de máxima verossimilhança, é ainda necessário considerar também todos os tópicos já discutidos na seção anterior. O denominador $\mathrm{L}(D)$ é uma integração sobre todas as possibilidades de topologias, tamanhos de ramo e valores para os parâmetros do modelo evolutivo, o que garante que a soma da probabilidade posterior para todos eles seja 1. O denominador atuará como um normalizador para o numerador. Reescrevendo, temos:

$$\mathrm{L}(filogenia|alinhamento) = \frac{\mathrm{L}(filogenia)\,\mathrm{L}(alinhamento|filogenia)}{\sum_{H} \mathrm{L}(filogenia)\,\mathrm{L}(alnhamento|filogenia)}$$

onde o termo filogenia descreve a topologia da árvore, o modelo de substituição e o comprimento dos ramos. Assim, através da multiplicação das probabilidades anteriores pela verossimilhança, divididos pelo fator de normalização, o método busca a hipótese (topologia da árvore, o modelo de substituição e o comprimento dos ramos) em que a probabilidade posterior é máxima.

O objetivo da inferência Bayesiana é calcular a probabilidade posterior para cada filogenia proposta. No entanto, para cada árvore diversos parâmetros devem ser especificados pelo usuário, incluindo topologia, tamanho dos ramos, parâmetros do modelo de substituição, parâmetros populacionais, relógio molecular, taxa evolutiva, etc. Dada uma filogenia, todos os parâmetros terão sua probabilidade posterior calculada. Se dadas 1000 filogenias, teremos 1000 valores de probabilidade posterior para cada parâmetro.

Devido à impossibilidade de construção de todas as filogenias possíveis para a maioria dos alinhamentos, a análise Bayesiana se aproveita de técnicas de amostragem para estimar os valores esperados de cada parâmetro.

Neste sentido, os métodos de inferência

Bayesiana utilizam as Cadeias de Markov Monte Carlo (MCMC, *Monte Carlo Markov Chain*) para aproximar as distribuições probabilísticas em uma grande variedade de contextos. Esta abordagem permite realizar amostragens a partir do conjunto total de filogenias, relacionando cada filogenia a um valor probabilístico. Sem a aplicação de um método que obtenha amostras do espaço de possíveis filogenias, como o modelo de MCMC, a estimativa de todos os parâmetros se tornaria analiticamente impossível nos atuais computadores.

Um dos métodos de MCMC mais usados na inferência filogenética é uma modificação do algoritmo Metropolis, chamado de Metropolis-Hastings. A ideia central deste método é causar pequenas mudanças em uma filogenia (topologia, tamanho dos ramos, parâmetros do modelo de substituição, etc.) e, após a modificação, aceitar ou rejeitar a nova hipótese de acordo com o cálculo de razão das probabilidades. Este método garante que diversas árvores sejam amostradas do espaço total de filogenias, amostrando filogenias com probabilidade posterior mais alta (Figura 15-5):

i) Inicialmente, o algoritmo MCMC gera uma filogenia aleatória X, arbitrariamente escolhendo o tamanho dos ramos para dar início à cadeia;

ii) O valor de probabilidade associado a esta filogenia é calculado (probabilidade posterior calculada através da fórmula de Bayes);

iii) Perturbações aleatórias são realizadas nesta filogenia inicial X (mudanças na topologia, no tamanho dos ramos, nos parâmetros do modelo de substituição, etc.) e geram uma filogenia Y;

iv) A probabilidade posterior é calculada para a filogenia Y;

v) A filogenia Y é tomada ou rejeitada para o próximo passo baseado na razão R (probabilidade posterior de Y dividida pela probabilidade posterior de X). Se R é maior que 1, a filogenia Y é tomada como base para o próximo passo. Se R é menor que 1, um número entre 0 e 1 é tomado aleatoriamente. Se R é maior que o número aleatório gerado, a filogenia será tomada, no entanto se for menor, a filogenia Y é rejeitada;

vi) Se a nova proposta Y for rejeitada, retorna-se ao estado X e novas modificações serão realizadas nesta filogenia;

vii) Supondo que a proposta Y tenha sido aceita, ela sofrerá uma nova perturbação a fim de gerar uma nova filogenia;

viii) Todas as árvores amostradas são armazenadas para posterior comparação. Os pontos visitados formam uma

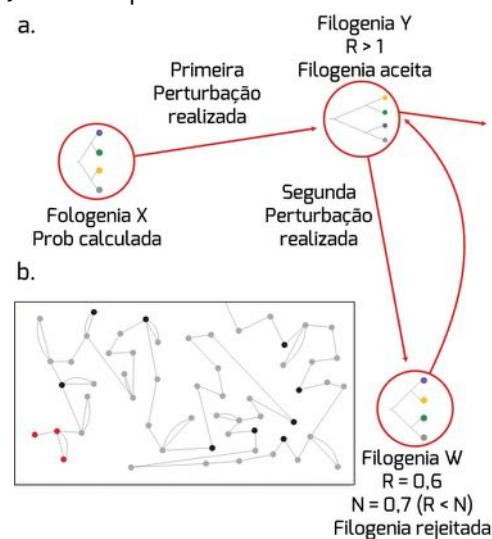



Figura 15-5: Esquema de amostragens MCMC aplicada à inferência filogenética pelo método Bayesiano utilizando o algoritmo de Metropolis-Hastings. (a) Após a proposição de uma filogenia inicial X, perturbações aleatórias são realizadas para gerar a filogenia Y. Devido à razão R>1, a nova filogenia é aceita. Nova perturbação é realizada para gerar a filogenia W e, devido a razão de probabilidades R resultar em um número menor que 1, um número aleatório N é sorteado. Sendo R<N, a nova proposição é rejeitada e a cadeia retorna à filogenia Y. (b) Andamento da cadeia na amostragem de filogenias. Cada círculo destaca uma nova filogenia que é proposta após a perturbação. As linhas conectando os círculos evidenciam a direção do andamento da cadeia. Apesar de a cadeia percorrer muitos passos, apenas alguns serão registrados para análise final (círculos pretos). Os círculos em vermelho são aqueles evidenciados em (a).

espécie de cadeia ao longo do espaço amostral total de filogenias.

O principal objetivo da cadeia é amostrar filogenias com probabilidades crescentes. No entanto, é importante que o algoritmo utilizado para tal permita que algumas árvores com menor probabilidade sejam amostradas para evitar que a cadeia fique "presa" em picos de máximo local (Figura 9-5).

Sendo assim, o cálculo da razão R considerando um valor aleatório entre 0 e 1 garantirá que, em determinados momentos, uma filogenia com menor probabilidade seja aceita. Por este método, é possível amostrar filogenias da região de um vale passando, por exemplo, de um pico de ótimo local para o pico de ótimo global (Figura 9-5).

A proposta de novas árvores na cadeia de Markov é uma etapa crucial para uma boa amostragem de filogenias. Na abordagem Bayesiana, uma boa amostragem inclui um grande número de filogenias, suficientemente diferentes entre si. Se filogenias muito diferentes são propostas, serão rejeitadas com muita frequência, pois é provável que tenham menor probabilidade posterior. Pelo contrário, se filogenias muito similares forem geradas, o espaço amostral não será varrido adequadamente e a cadeia deverá "correr" por muitos passos (amostrar um maior número de filogenias), aumentando o tamanho da cadeia e o tempo computacional.

Estimar o quanto a cadeia deve percorrer para amostrar um número suficiente de filogenias para as sequências dadas (espaço de árvores) é um fator fundamental para obter bons resultados em uma análise Bayesiana. Na maioria dos programas que utilizam estatística Bayesiana para inferir filogenias, o usuário deve especificar o tamanho da cadeia. Esse número é de grande subjetividade, e depende diretamente da distribuição das probabilidades anteriores, do número de táxons incluídos na filogenia e da relação evolutiva entre eles.

A Figura 16-5 exemplifica o andamento da amostragem da MCMC em um espaço de filogenias. Supondo que os quadrados em a, b e c representam um espaço amostral de filogenias, semelhante ao apresentado na Figura 15-5b, e que os pontos pretos sejam as filogenias que vão sendo amostradas com o desenvolvimento da MCMC vemos que, ao final do processo, depois de empregados 100 mil passos (Figura 16-5c), um grande número de filogenias foi amostrado.

Ainda, na região delimitada por um círculo, assumimos que estão as filogenias com maior probabilidade de explicar a história evolutiva de um grupo de organismos, ou seja, as filogenias reais. Note que quanto maior o número de passos percorridos pela cadeia, maior a amostragem do espaço de filogenias e maior o número de amostras dentro da região com filogenias de alta probabilidade.

Ao final, após o término da cadeia, a distribuição das probabilidades posteriores de todos os parâmetros deve ser verificada. No

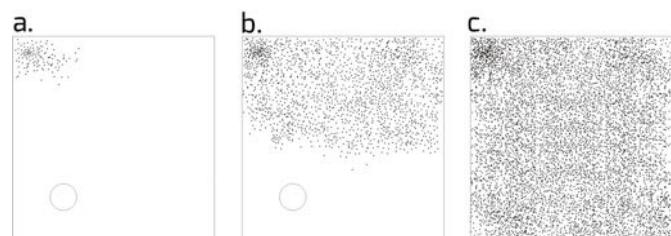



Figura 16-5: Espaço de possíveis árvores analisadas pela MCMC. Considerando que os quadrados descrevem o espaço amostral de todas as filogenias possíveis para um dado conjunto de sequencias, os pontos pretos representam as filogenias que foram amostradas ao longo da cadeia. Os círculos presentes no canto esquerdo inferior representam a região de máximo global (isto é, maior probabilidade) neste espaço amostral. O andamento da cadeia neste exemplo é o mesmo apresentado na Figura 15-5b (a) cento e trinta passos percorridos pela cadeia; (b) trinta mil passos percorridos pela cadeia; (c) cem mil passos percorridos pela cadeia. Nota-se que quanto maior o número de passos percorridos, maior a amostragem de filogenias no espaço. Da mesma forma, aumenta a probabilidade de a cadeia amostrar aquelas filogenias de máximo global.

entanto, as amostras tomadas no início da cadeia são tipicamente descartadas, pois estão sob forte influência do local de início da cadeia. As filogenias do início da cadeia estão muito longe de pontos máximos no espaço amostral e, por isso, é provável que todas as novas filogenias sugeridas subsequentemente sejam tomadas para o próximo passo (qualquer árvore proposta será mais provável que as árvores iniciais semelhantes àquela gerada aleatoriamente).

Esta fase inicial é conhecida como período de *burn in* (Figura 17-5). Conforme a cadeia avança, espera-se que a probabilidade das árvores amostradas aumente e, quando um número suficiente de filogenias for amostrado, chegue a uma distribuição estacionária. Em termos Bayesianos, espera-se que a cadeia atinja a convergência.

Um dos primeiros indicativos de que a cadeia convergiu para a distribuição correta está na estabilidade dos valores de probabilidade dos parâmetros da cadeia (cada parâmetro da filogenia poderá ter uma distribuição independente). Portanto, a representação gráfica dos valores das probabilidades e dos respectivos passos da cadeia (*trace plot*) é uma importante ferramenta para monitorar o desempenho da MCMC (Figura 17-5).

Devido ao aumento brusco de probabilidade das filogenias que são visitadas pelo andamento da cadeia, os gráficos necessariamente incluirão os valores medidos em escala logarítmica ($\ln L$, Figura 17-5). Em estatística Bayesiana, é comum que seja atribuído um intervalo de credibilidade de 95% para os parâmetros amostrados. Estes valores são obtidos através da eliminação de 2,5% dos valores mais baixos e de 2,5% dos valores mais altos para um determinado parâmetro. Um intervalo de credibilidade contém o valor correto com 95% de probabilidade; no entanto, não se trata de um intervalo de confiança.

Adicionalmente, outros métodos são úteis para diagnosticar a convergência da cadeia, tais como o exame do tamanho amostral efetivo (ESS) e a comparação de amostras resultantes de diferentes cadeias (várias cadeias de MCMC são aplicadas para o mesmo conjunto de dados). Apesar de ser computacionalmente intensiva, a última alternativa parece ser a mais confiável para verificar a convergência. Contudo, o exame de ESS é, ainda hoje, o método mais utilizado. O tamanho amostral efetivo é uma estimativa para verificar o número de amostras independentes existentes na cadeia, ou seja, quantas amostras não similares foram tomadas. Atualmente, um ESS maior que 200 é um indicativo de que a cadeia convergiu adequadamente.

A técnica de *Metropolis Coupling*, conhecida como MCMCMC ou $(MC)^3$, através da introdução da corrida simultânea de duas cadeias, pode ajudar na amostragem de máximos globais e beneficiar na convergência da cadeia. Nesta técnica uma cadeia, chamada de quente (*hot chain*), permite aproximar os valores de máxima e mínima probabilidade das amostras para que a cadeia possa, de forma mais rápida, “saltar” entre picos de probabilidade, especialmente de máximos locais para máximos globais. O aquecimento da cadeia é dado pelo parâmetro β e visa diminuir a altura dos picos locais no espaço amostral. Uma segunda cadeia simultânea, chamada de fria (*cold chain*), utiliza as informações destes saltos da cadeia quente para melhorar a sua

Figura 17-5: Representação gráfica das probabilidades das filogenias na cadeia ao longo de 300 milhões de amostragens. O esquema demonstra duas visualizações possíveis: à esquerda, são mostrados apenas os pontos referentes às amostras tomadas ao longo da cadeia e, à direita, as amostragens sucessivas são ligadas umas as outras para facilitar a visualização do comportamento da cadeia. Em cinza, a fase inicial de *burn in* da Cadeia de Markov Monte Carlo.

amostragem e garantir a convergência.

Os métodos Bayesianos de inferência filogenética ainda têm a vantagem de aplicar modelos que envolvem diferentes tipos de relógios moleculares.

As distâncias genéticas, depois de "tratadas" pelos modelos de substituição, não tem qualquer significado sozinhas quando se deseja estimar, por exemplo, a idade do ancestral comum mais recente de duas OTUs. Esta e outras questões podem ser avaliadas quando aplicamos uma medida de tempo nas inferências, a fim de calibrar as taxas evolutivas. Sequenciamentos de amostras isoladas em diferentes épocas podem fornecer a calibração adequada para inferências temporais, pois se assume uma taxa evolutiva constante ao longo de um tempo *t* para todos os ramos de uma filogenia (relógio molecular estrito).

As taxas evolutivas dependem de diversos fatores e podem variar, nem sempre seguindo a constância proposta por este modelo. Após a introdução de um tipo específico de relógio molecular relaxado, as taxas de evolução podem variar ao longo da árvore para diferentes grupos e não são correlacionadas, ou seja, grupos evolutivamente próximos não necessariamente terão taxas de evolução semelhantes (relógio molecular relaxado não correlacionado).

Complexos modelos de dinâmica populacional podem ser analisados sob uma perspectiva Bayesiana. Quando o conjunto de sequências submetido às analises são isolados de uma população homogênea, os parâmetros de história demográfica podem ser usados para modelar as mudanças populacionais ao longo do tempo. Desta forma, através da estatística Bayesiana é possível, além da inferência filogenética, refinar as análises e datar filogenias e ramos específicos (Figura 18-5), inferir caracteres ancestrais e analisar a dinâmica populacional sob uma ótica evolutiva.

Figura 18-5: Árvore filogenética consenso gerada por inferência Bayesiana para 70 sequências de nucleotídeos. As cores nos ramos representam diferentes clados (B-H). O grupo externo está identificado como clado A. O Clado H foi agrupado para facilitar a representação. Nos nós estão especificados os valores de probabilidade posterior acima de 70. Abaixo, é apresentada a escala temporal inferida a partir da utilização de um relógio molecular relaxado.

5.8. Confiabilidade

O papel principal das técnicas de inferência filogenética é desvendar as relações evolutivas reais através de dados moleculares, buscando garantir que esta reconstrução seja fidedigna. Além da inferência das relações evolutivas entre os táxons, é igualmente importante que a filogenia possua precisão. Esta característica está relacionada ao número de filogenias que podem ser excluídas, a partir do conjunto total de filogenias, por não serem "verdadeiras". Quanto maior o número de filogenias excluídas neste processo, mais preciso é o método.

Em geral, na maioria dos casos de reconstrução filogenética, a falta de precisão das filogenias está relacionada ao conjunto de dados que está sendo fornecido no alinha-

mento. O gene considerado, o tamanho das sequências, o número de indivíduos e o grupo externo são atribuições fundamentais para uma reconstrução filogenética precisa e dependem, especialmente, do objetivo do estudo e da própria disponibilidade de informação.

Em muitos casos, o pesquisador é ainda dependente do número de amostras e do sucesso de coleta em campo, sobretudo, quando seu objeto de estudo se trata de uma espécie rara ou de indivíduos de difícil amostragem. No entanto, apesar de toda a informação relacionada ao conjunto de dados, a dificuldade de amostragem de indivíduos parece ser, sem dúvida, o principal problema relacionado a precisão das filogenias, pois a falta de dados de variabilidade genética compromete a inferência de história evolutiva coerente.

Como é possível saber se a amostragem foi suficiente e a filogenia é confiável? Usualmente, a resposta para esta questão consiste na reamostragem de dados. Se novas amostras forem tomadas e a mesma filogenia for reproduzida, a filogenia proposta tem seu valor reforçado. No entanto, na maioria dos casos, a reamostragem de dados da forma usual (coletas de novos espécimes, reamostragens em campo, achado fóssil diferente, etc) não é factível. Assim, algoritmos que produzem diferentes amostragens utilizando o mesmo conjunto de dados foram desenvolvidos para possibilitar a verificação da confiabilidade nos clados das filogenias. Destaca-se entre estes algoritmos o método de *bootstrap*.

Bootstrap é um método de reamostragem utilizado para realizar comparações da variabilidade das hipóteses filogenéticas, oferecendo medidas de confiabilidade aos clados propostos. A reamostragem é realizada a partir do mesmo conjunto de dados, e novas amostras fictícias com o mesmo tamanho serão geradas.

Segundo este método, cada sítio do alinhamento será tratado de forma independente. Conforme a Figura 19-5, inicialmente o algoritmo reconstruirá a filogenia a partir do alinhamento dado e, posteriormente, diversas replicatas serão reconstruídas. As colunas, representando os sítios do alinhamento, serão aleatoriamente tomadas (amostradas) pelo algoritmo e, em seguida, serão agrupadas uma ao lado da outra de maneira a formar um novo alinhamento (com o mesmo número de sítios do alinhamento original, Figura 19-5).

Por este método, é possível que um mesmo sítio seja amostrado mais de uma vez e, portanto, alguns sítios não serão selecionados para o novo alinhamento. Um número fornecido pelo usuário especificará o número de pseudoreplicatas (novos alinhamentos) que serão construídas. Assim que uma pseudoreplicata for criada, o algoritmo constrói a filogenia correspondente.

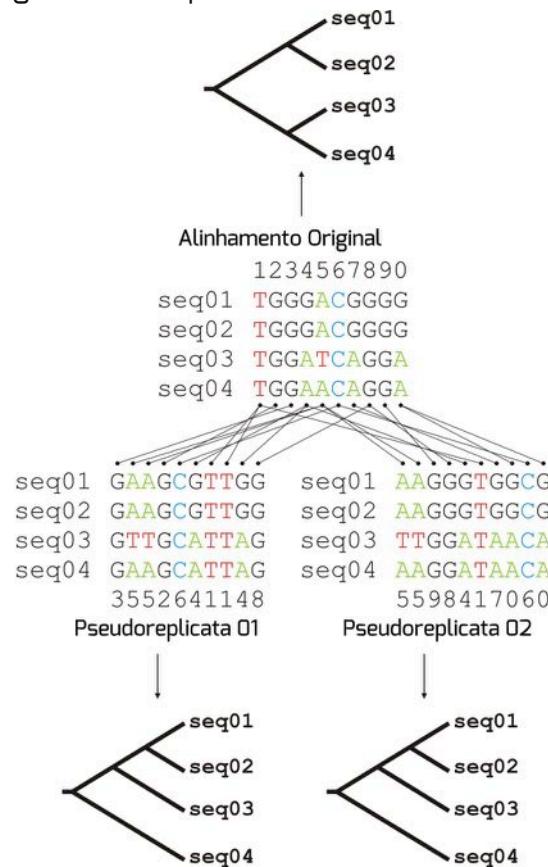



Figura 19-5: Método de *boostrap* para filogenias. A partir do alinhamento original, as colunas que representam os sítios serão aleatoriamente amostradas para construir pseudoreplicatas (um mesmo sítio pode ser sorteado diversas vezes). Estas, por sua vez, serão utilizadas para a inferência de filogenias, da mesma forma que o alinhamento original.

É importante ressaltar que a inferência destas filogenias será realizada pelo método de construção especificado pelo usuário, seja aproximação de vizinhos, máxima parcimônia ou máxima verossimilhança (para árvores bayesianas, veja adiante). Ao final, o algoritmo analisará os clados e automaticamente verificará a presença de determinados agrupamentos em todas as filogenias construídas. Se, por exemplo, encontramos as sequências 1 e 2 formando um clado em 70% das filogenias construídas, atribuiremos a confiabilidade de 70 ao clado formado por estas duas sequências. Comumente, o valor de confiabilidade dos clados é colocado próximo ao ancestral comum do clado (Figura 18-5).

A partir dos resultados de confiabilidade dos clados é possível também construir filogenias baseando-se na árvore consenso gerada pela regra da maioria (*majority-rule consensus tree*). Neste método, o algoritmo tabulará todos os clados formados em todas as replicatas geradas. Aqueles clados que mais aparecerem servirão para montar a filogenia consenso.

Ao contrário dos métodos de aproximação de vizinhos, máxima parcimônia e máxima verossimilhança, a confiabilidade de filogenias construídas através de estatística Bayesiana é inerente ao processo. Como diversas filogenias são amostradas ao longo do desempenho da Cadeia de Markov, não é necessário nenhum método para simular reamostragens do mesmo conjunto de dados. As amostras serão resumidas a partir da distribuição posterior de filogenias como frequência de clados individuais e serão identificadas por um número próximo ao ancestral comum daqueles clados (Figura 18-5). Portanto, o valor de probabilidade posterior de um clado representa uma inferência a respeito da probabilidade daquele clado.

A comparação dos valores de *bootstrap* e de probabilidade posterior dos clados para filogenias construídas a partir do mesmo alinhamento utilizando máxima verossimilhança e o método Bayesiano, respectivamente, leva a conclusão de que o método Bayesiano superestima a confiança aos clados. A confiança atribuída pela probabilidade posterior é geralmente maior que aquela atribuída pelo método de *bootstrap*. Por isso, enquanto uma confiança acima de 70 é considerada sustentada para o *bootstrap*, apenas valores acima de 90 podem ser considerados relevantes para os métodos Bayesianos.

5.9. Interpretação de filogenias

Árvores filogenéticas são diagramas que denotam a história evolutiva de diferentes OTUs a partir de seu ancestral comum. Mais do que isso, as filogenias moleculares são ferramentas que ajudam no entendimento dos diversos processos evolutivos que moldam o genoma dos organismos. Desta forma, a interpretação das implicações evolutivas associadas a um, ou a um conjunto de táxons, está diretamente relacionada à disposição dos ramos internos e externos de uma árvore. Independentemente do método de inferência, ou da forma como a árvore é apresentada, a interpretação dos resultados será baseada nos mesmos pressupostos, ainda que métodos diferentes possam originar filogenias diferentes.

Inicialmente, é necessário observar a presença de uma raiz. Como já discutido, o método de enraizamento pelo grupo externo é o mais comum e utiliza organismos sabidamente relacionados ao grupo em evidência, servindo para orientar o algoritmo em relação às características mais ancestrais do grupo. O grupo externo ajudará a evidenciar o tempo evolutivo. Na Figura 20-5, por exemplo, o grupo externo é dado pelo orangotango, pois este compartilha o mesmo ancestral comum que o restante do grupo. No caso de filogenias sem raiz, é necessário ter cautela nas interpretações, pois este tipo de diagrama apenas revela a relação entre os táxons.

Depois de encontrada a raiz da filogenia, é preciso avaliar os ramos. Dependendo do método, os ramos podem ter significados diferentes. Na Figura 18-5, os ramos evidenciam o tempo real, apresentando OTUs amostradas no passado. Pelo contrário, na Figura 20-5, os ramos evidenciam apenas um

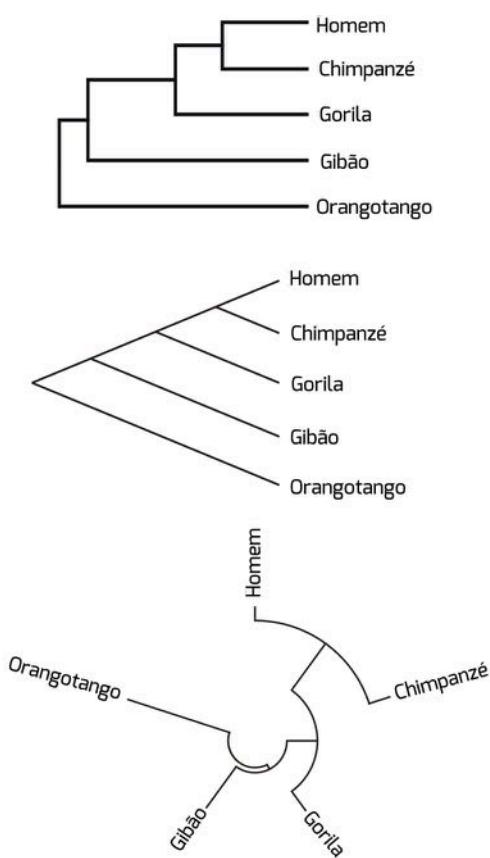



Figura 20-5: Diferentes representações da filogenia dos primatas.

tempo evolutivo representado pelo número de modificações genômicas, desde o organismo ancestral até os ramos terminais. Além disso, deve-se perceber a escala na qual os ramos foram representados, pois estes indicam o número de substituições que provavelmente ocorreram ao longo do processo evolutivo e podem ajudar na interpretação das taxas evolutivas.

Conclusões evolutivas baseadas em árvores filogenéticas devem ser sustentadas em árvore confiáveis e, por isso, a medida de confiabilidade dos ramos deve ser denotada. Inicialmente, é necessário verificar o método utilizado para reconstrução da filogenia e, quando necessário, verificar o algoritmo utilizado para gerar a confiabilidade dos clados. Ramos com maiores valores de confiabilidade gerarão conclusões mais confiáveis, enquanto que clados com baixos valores deverão ser interpretados com maior cuidado. No entanto, não é necessário negar totalmente conclusões baseadas em filogenias com baixa confiabilidade nos ramos. O tipo de método, a forma de amostragem e o número de OTUs podem ser fatores de interferência e, assim, podem prejudicar a valorização dos ramos.

O padrão de organização dos ramos de uma filogenia denota o padrão de ancestralidade. As filogenias não são escadas, onde alguns organismos são “mais evoluídos” que outros, mas uma representação da história da derivação de OTUs. Na Figura 18-5, por exemplo, é possível observar que os clados B, C, D, E, F e G possuem um ancestral comum que compartilha um outro ancestral com o clado A. Já o clado H, representado por um triângulo para evidenciar um grande número de táxons naquele ponto da filogenia, teve um ancestral comum dentro do clado G. Este padrão sugere que o clado H se originou a partir do clado G. Da mesma forma, podemos observar a disposição do clado G em relação ao F e concluir que o primeiro se originou a partir do segundo.

No caso da Figura 20-5, observamos que humanos e chimpanzés tiveram um mesmo ancestral comum. Com base nestes dados, é incorreto pensarmos que humanos são derivados de chimpanzés, ou que humanos são mais evoluídos que chimpanzés. Estes organismos estão apenas formando um mesmo clado dentro da filogenia dos primatas.

Por último, é fundamental saber o objetivo do estudo filogenético a ser realizado. Árvores filogenéticas devem ser construídas para responder uma determinada questão, que pode envolver apenas um, ou diversos organismos.

Quando possível, é importante reconstruir a filogenia utilizando diferentes métodos de inferência e compará-las entre si. A conclusão desta forma será melhor sustentada. Além disso, atualmente, a história retratada em uma filogenia não é por si só satisfatória. Outras ferramentas podem ser utilizadas para complementar e sustentar a interpretação de uma filogenia, incluindo análises de recombinação, pressão seletiva e estruturação populacional, verificação de coespeciação, construção de redes filogeográficas, compa-

ração com dados de fósseis, eventos geológicos, dados históricos e, até mesmo, análises de dados comportamentais.

Um exemplo da combinação de análises filogenéticas com dados históricos veio na confirmação da origem e disseminação humana a partir da África. Através da utilização de dados histórico-antropológicos (como vestígios materiais de hominídeos ancestrais), fósseis de hominídeos e análises de DNA mitocondrial de representantes de diferentes etnias, os pesquisadores puderam traçar as rotas de disseminação humana a partir da África.

Outro exemplo está na solução de um enigma que perturbou zoólogos por um longo período: a posição taxônomica do panda-gigante entre os mamíferos carnívoros. Apesar de esta espécie ser fisicamente muito similar a um urso, outras características, como dentição e anatomia das patas, levaram à proposição de uma hipótese antes não imaginada.

Tal hipótese propunha que o panda-gigante (*Ailuropoda melanoleuca*) seria proximamente relacionado ao o panda-vermelho (*Ailurus fulgens*), um mamífero de pequeno porte, semelhante ao guaxinim. Com o emprego de diferentes dados, incluindo fósseis, anatomia de mamíferos atuais, distribuição geográfica, sequências de DNA de diferentes porções do genoma, sequências de aminoácidos de diferentes proteínas e mapeamento cromossômico, foi possível estabelecer uma história evolutiva plausível, capaz de descrever a origem evolutiva do panda-gigante (Figura 21-5).

Por meio dessa análise combinada de dados, se propôs que o panda-gigante, um urso, derivou do ancestral comum dos ursos há cerca de 24 milhões de anos, muito antes das derivações que originaram todos os outros ursos existentes hoje. Além disso, observou-se que os ursos e os procionídeos (grupo que inclui o guaxinim e o panda-vermelho) possuem um ancestral comum que deu origem às duas linhagens há aproximadamente 30 milhões de anos.

A filogenia molecular é uma ferramenta útil quando empregada isoladamente, mas que pode se beneficiar de diferentes tipos de dados para propor uma história evolutiva. Em última análise, a decisão sobre que tipos de

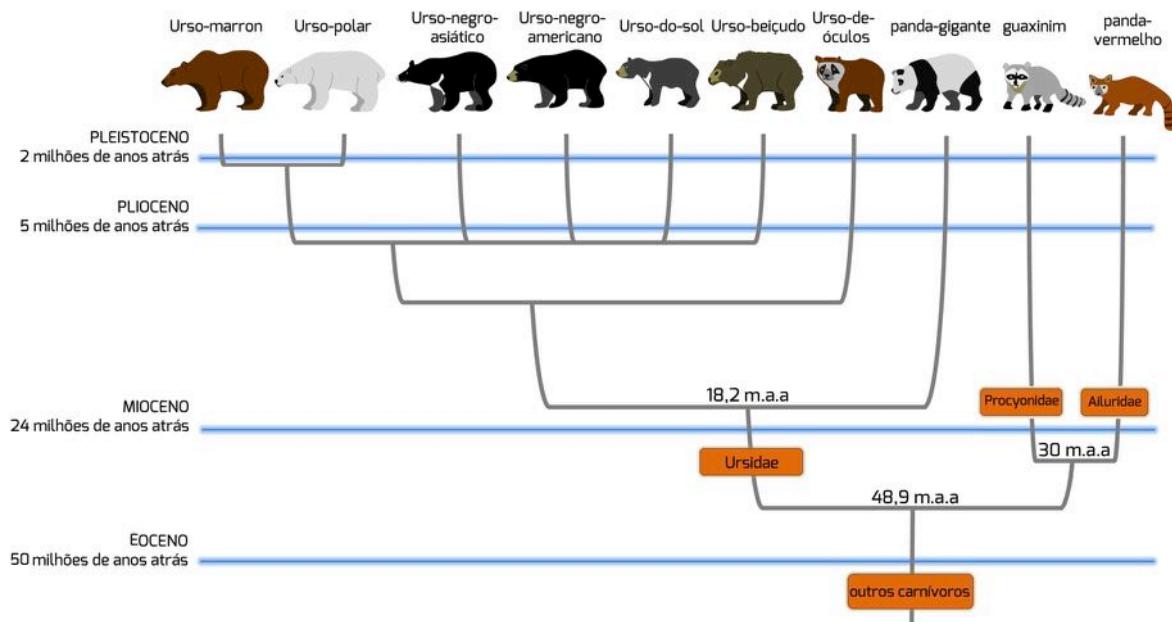



Figura 21-5: Posição filogenética do panda-gigante, baseada na combinação de diferentes tipos de dados. Baseado em BININDA-EMONDS, Olaf R.P. *Phylogenetic position of the giant panda*. Em: LINDBURG, D.G. & Baragona, K. *Giant pandas: Biology and conservation*. Berkeley: University of California Press, 2004; e em EIZIRIK, Eduardo e colaboradores: *Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences*. *Mol Phylogenet Evol*, *56*, 49, 2010.

dados (além dos moleculares) serão empregados na análise filogenética dependerá da pergunta a ser respondida com essa técnica. Não existem regras pré-estabelecidas, e as estratégias analíticas precisam ser propostas caso a caso.

5.10. Conceitos-chave

Ancestral: organismo ou sequência que originou novo(s) organismo(s) ou sequência(s). Em alguns casos pode ser considerado o mesmo que primitivo.

Apomórfico: refere-se a um caractere novo adquirido ao longo do processo evolutivo, uma inovação. Uma apomorfia pode servir de diagnóstico para separação de clados.

Aproximação dos vizinhos: *neighbor joining* (NJ), método de inferência filogenética quantitativo baseado em distância genética.

Autapomorfias: apomorfias específicas e restritas a um clado.

Bootstrap: método de reamostragem que permite verificar a confiabilidade dos ramos de uma filogenia.

Cadeias de Markov Monte Carlo: método utilizado pela estatística Bayesiana para amostrar as probabilidades de distribuição de diferentes parâmetros das filogenias.

Clado: grupo formado por um ancestral e todos seus descendentes, um ramo único em uma árvore filogenética.

Derivado: que se originou de um ancestral e é mais recente no tempo evolutivo (nota: deve-se evitar o termo "mais evoluído" e, em seu lugar, empregar "derivado").

Distância Genética: medida quantitativa da divergência genética entre organismos.

Espaço Amostral de Filogenias: espaço teórico que inclui todas as filogenias possíveis (com raiz ou sem raiz) para um determinado alinhamento.

Frequência de equilíbrio: ponto em que não existe mais alteração nas frequências dos alelos.

Grupos irmãos: clados que dividem um ancestral comum.

Homologia: similaridade originada por ancestralidade comum.

Inferência filogenética Bayesiana: método qualitativo de inferência filogenética baseado na estatística Bayesiana. Através da Cadeia de Markov Monte Carlo este método buscará as árvores mais prováveis dentro das filogenias amostradas.

Máxima Parcimônia: método qualitativo de inferência filogenética que busca a árvore que minimiza o número total de substituição de nucleotídeos.

Máxima Verossimilhança: método qualitativo de inferência filogenética que busca a árvore com a máxima verossimilhança.

Monofilia: associação entre o ancestral comum e todos os seus descendentes, formando um clado monofilético.

Múltiplas Substituições: eventos múltiplos de substituição de nucleotídeo localizado em um mesmo sítio do DNA.

Modelos de Substituição: modelos matemáticos utilizados para descrever o processo evolutivo ao longo do tempo, podendo ser aplicados ao alinhamento de nucleotídeos ou aminoácidos.

Ortólogo: genes homólogos em diferentes organismos e que mantém a mesma função.

OTU: unidade taxonômica operacional, folha ou nó terminal em uma árvore filogenética.

Parafilia: associação entre o ancestral comum e apenas parte de seus descendentes, formando um clado parafilético.

Parálogo: genes homólogos de um mesmo organismo que divergiram após duplicação.

Plesiomórfico: dotado de características do ancestral que são conservadas nos descendentes.

Polifilia: associação entre diferentes OTUs sem a necessidade de um único ancestral comum, frequentemente originada por convergência evolutiva.

Primitivo: diz-se de características ou organismos ancestrais, anteriores no tempo evolutivo a organismos ou características mais recentes.

Probabilidades Anteriores: distribuição dos valores de um parâmetro filogenético que é sabido de antemão pelo pesquisador.

Probabilidades Posteriores: conjunto da distribuição dos valores de parâmetros filogenéticos resultantes do método de inferência Bayesiana.

Sistemática: estudo da diversificação das formas vivas e suas relações ao longo do tempo.

Taxonomia: estudo que busca agrupar os organismos com base em suas características e nomear os grupos obtidos, classificando-os em alguma escala.

Taxon: grupo (de qualquer nível hierárquico) proposto pela taxonomia.

Topologia: descreve a ordem e a disposição exata das OTUs em uma filogenia.

UPGMA: *unweighted pair-group method using arithmetic average*, método de inferência filogenética quantitativo baseado em distância.

5.11. Leitura recomendada

FELSENSTEIN, Joseph. ***Inferring Phylogenies***. Sunderland: Sinauer, 2004.

GREGORY, T. Ryan: ***Understanding Evolutionary Trees***. Evo. Edu. Outreach, 2008, 1,121-137.

LEMEY, Philippe; SALEMI, Marco; Vandamme, Anne-Mieke (Org.). ***The Phylogenetic Handbook***. 2.ed. Cambridge: Cambridge University Press, 2009.

MATIOLI, Sergio Russo; FERNANDES, Flora M.C. (Org.). ***Biologia Molecular e Evolução***. 2.ed. Ribeirão Preto: Holos, 2012.

NEI, Masatoshi; KUMAR, Sudhir. ***Molecular Evolution and Phylogenetics***. Nova Iorque: Oxford University Press, 2000.

PABÓN-MORA, Natalia; GONZÁLEZ, Favio. A classificação biológica: de espécies a genes. In: ABRANTES, Paulo C. (Org.), ***Filosofia da Biologia***. Porto Alegre: Artmed, 2011.

SCHNEIDER, Horacio. ***Métodos de Análise Filogenética: Um Guia Prático***. 3.ed. Ribeirão Preto: Holos, 2007.

6. Biologia de Sistemas

"Pensar a complexidade – esse é o maior desafio do pensamento contemporâneo, que necessita de uma reforma no nosso modo de pensar."

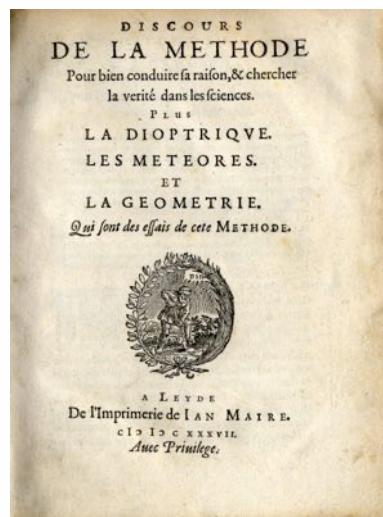
Edgar Morin & Jean-Louis Le Moigne

Joice de Faria Poloni
Bruno César Feltes
Fernanda Rabaioli da Silva
Diego Bonatto

6.1. Introdução

Uma das posturas metodológicas mais significativas do pensamento científico contemporâneo consiste em reduzir o todo a suas partes componentes. Por exemplo, entendemos o funcionamento de um organismo como fruto da ação de órgãos. Estes por sua vez, são compostos por tecidos, que são compostos por células. As células têm como componentes moléculas que, por fim, são compostas por átomos.

DISCOURS
DE LA METHODE
Pour bien conduire sa raison, & chercher la verité dans les sciences.
PLUS
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui sont des essais de cete METHODE.
A LEYDE
De l'Imprimerie de IAN MAIRE.
CIↃ IↃ C XXXVII.
Avec Privilege.

Esta abordagem, especialmente importante e difundida na área biológica, é fruto das idéias introduzidas pelo filósofo René Descartes em meados do século XVII, indicando que cada problema encontrado deve ser dividido em tantas pequenas partes quanto for necessário para resolvê-lo de maneira mais parcimoniosa.

É neste contexto que emerge a divisão disciplinar no estudo da natureza. Desde os tempos da escola até a universidade, o conhecimento a ser ensinado manifesta-se na separação das disciplinas. Por exemplo, no meio acadêmico observamos a biologia compartimentada em botânica, zoologia, ecologia, genética, biologia celular e essas, por sua vez, subdivididas em outras áreas. Como aspecto positivo, o estudo das partes forma especialistas e divide o trabalho, facilitando o entendimento de suas partes componentes. Contudo, neste processo tem-se uma redução da complexidade característica dos fenômenos naturais, o que pode comprometer nossa capacidade de entendê-los.

De fato, a complexidade é inerente à biologia, ao funcionamento do nosso organismo e à natureza. Há a necessidade, assim, da construção de uma abordagem que inclua esta complexidade, de forma sistêmica; que interligue as diversas interações presentes e que, ao confrontá-las, consiga encontrar relações mais informativas e completas.

A partir desta premissa, emergem na década de 1950 as primeiras concepções sobre a Biologia de Sistemas (BS). Essa área, pautada nos conceitos de sistema e de complexidade, envolve um estudo sistemático de interações em um sistema biológico.

O conceito de sistema é entendido como um conjunto de partes ou elementos que possuem relações entre si, relações estas

que diferem-se daquelas realizadas com outros elementos, fora do sistema. Já a idéia de complexidade é definida como a condição de elementos de um sistema e a relação entre esses elementos em um determinado momento.

Um sistema complexo, por conseguinte, é um sistema composto de partes interconectadas que, como um todo, exibe uma ou mais propriedades que não seriam observadas a partir das propriedades dos componentes individuais, possibilitando assim a observação de novos fenômenos. Portanto, a BS é um campo que investiga as interações entre os componentes de um sistema biológico, buscando contribuir no entendimento de como estas interações influenciam a função e o comportamento do sistema.

A busca da compreensão da biologia em nível de sistema é um tema recorrente na comunidade científica. Norbert Wiener, em 1948, foi um dos proponentes da abordagem sistemática que levou ao nascimento da cibernética, ou biocibernética, consolidada com os estudos do médico neurologista, William Ross Ashby (1903-1972). A partir de 1959, Robert Rosen, sob orientação do professor Nicolas Rashevsky, propôs uma metodologia baseada na "biologia relacional", onde o mais importante na biologia era o estudo da vida em si. Após 20 anos, Ludwig von Bertalanffy (1901-1972) criou a teoria geral dos sistemas, tornando-se o precursor da BS. Em 1966 foi formalizado o estudo da BS, com o lançamento da disciplina "Teoria e Biologia de Sistemas" pelo teórico de sistemas Mihajlo Mesarovic (1928).

A partir do trabalho destes pesquisadores, a teoria geral dos sistemas pode ser definida como a área que estuda a organização abstrata de fenômenos, investigando todos os princípios comuns a todas as entidades complexas (não somente biológicas) e os modelos que podem ser utilizados para a sua descrição.

Com o avanço da biologia molecular nas décadas que se seguiram, juntamente com o nascimento da genômica funcional, grandes quantidades de dados tornaram-se disponíveis e os bancos de dados e ferramentas de análise adaptaram-se ao volume crescente de informações, permitindo construir modelos mais amplos, capazes de lidar com aspectos e fenômenos inacessíveis até então. Assim em 2000, quando o Instituto de Biologia de Sistemas foi fundado, a biologia de sistemas emergiu como um campo próprio, estimulado pelo aumento de dados "ômicos" e pelos avanços da parte experimental e da bioinformática visando o entendimento sistemático da biologia. Desde então, grupos de pesquisas dedicados à BS têm sido formados em todo o mundo.

Para tal, a BS depende de ferramentas interdisciplinares para obter, integrar e analisar diversos tipos de dados, exemplificados na Tabela 1-6. Essa abordagem requer novas técnicas de análise, ferramentas de informática, métodos experimentais e uma nova postura metodológica, articulando partes normalmente estudadas separadamente.

6.2. Biologia de Sistemas

Em suas análises, a BS relaciona partes individuais de um sistema como representações gráficas de conjuntos de nós ou vértices (V), conectados entre si por conectores ou arestas (E, do inglês *edge*). Os nós podem representar indivíduos, proteínas ou mesmo lugares, enquanto que os conectores representam a conexão que está presente entre cada par de nós. Esta representação gráfica é denominada de rede.

Muitos exemplos de rede podem ser citados, como redes de cadeia alimentar, amplamente aplicadas na ecologia, redes neurais e de interação proteica usadas na biologia e ciências médicas, além da própria *World Wide Web*, que representa uma das maiores redes funcionais no mundo da comunicação e informática.

A análise matemática de redes é denominada de teoria de grafos, e consiste em um dos principais objetos de estudo da matemática discreta. Desta forma, o termo "rede" representa as interações funcionais de um sistema, enquanto que o termo "grafo" enfa-

Tabela 1-6: Ferramentas utilizadas no estudo da BS.

Área	Tipo de análise
Bioinformática	Funções biológicas por meio de ferramentas da informática
Genômica	Sequências de DNA
Transcriptômica	Transcritos
Proteômica	Proteínas
Interatômica	Interações proteicas
Interferômica/ microRNômica	RNAi/miRNA
Epigenômica	Modificações na cromatina e no DNA
Metabolômica	Metabólitos
Fluxômica	Alterações dinâmicas de moléculas dentro de uma célula ao longo do tempo
Biômica	Bioma
Glicômica	Totalidade de carboidratos
Farmacogenômica	Genes que definem o comportamento da droga
Nutrigenômica	Relação entre a dieta e os genes individuais
Toxicogenômica	Estrutura e atividade do genoma e os efeitos biológicos adversos na exposição a xenobióticos
Imunômica	Função molecular associada aos transcritos de RNAm relacionados à resposta imune

tiza as análises matemáticas deste sistema. Neste capítulo, contudo, usaremos ambos os termos como sinônimos.

Historicamente, a teoria de grafos foi desenvolvida em 1736 pelo matemático suíço Leonard Euler na resolução do problema das sete pontes de Königsberg, atualmente conhecida como Kaliningrado, na Rússia. A cidade de Königsberg é atravessada pelo Rio Pregel e consiste de duas grandes ilhas que eram conectadas entre si e com as margens opostas por sete pontes (Figura 1A-6). O problema apresentado a Euler consistia em descobrir como caminhar pela cidade atravessando cada ponte apenas uma vez. A técnica desenvolvida pelo matemático suíço foi adaptar o mapa de Königsberg, transformando as margens e ilhas em nós e as pontes em conectores (Figura 1B-6). Euler submeteu a rede que desenvolveu a análises matemáticas, porém não encontrou solução para o problema. Contudo, a metodologia de análise de Euler foi um marco histórico na análise de problemas combinatórios, além de estabelecer o conceito de topologia que é usado em BS (ver adiante).

O emprego da teoria de grafos e suas aplicações têm apresentado um crescimento explosivo devido a sua multidisciplinaridade e ao seu conceito de modelo que permite estudar um objeto específico sem negligenciar o meio em que este objeto se encontra. Por exemplo, é possível estudar determinado fármaco considerando a atividade que diversos compostos e enzimas poderiam exercer sobre ele. Nesses estudos pode-se construir uma rede onde os nós representam compostos e enzimas e os conectores representam se há ou não relação entre eles, permitindo analisar:

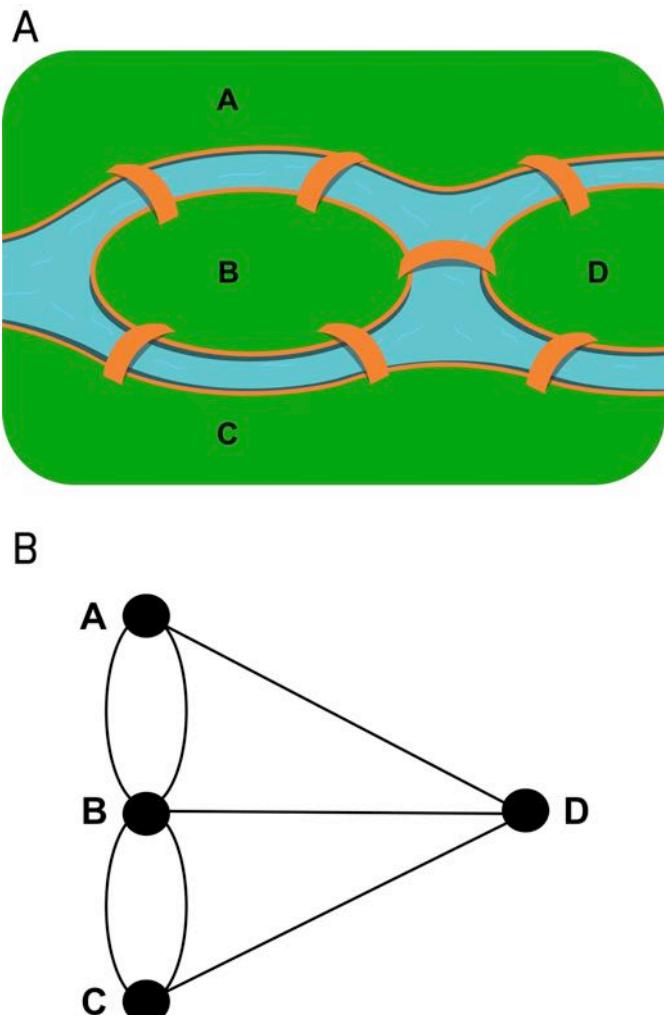



Figura 1-6: (A) Representação parcial do mapa de Königsberg e suas setes pontes. (B) Ilustração da rede desenvolvida por Euler.

i) a conectividade dos compostos ou enzimas, ou seja, que tipo de relação duas moléculas aleatórias podem apresentar na rede;
ii) a centralidade, que caracteriza as moléculas que apresentam maior influência sob a ação do fármaco em questão.

Conceitos básicos de grafos

Considerando-se a estreita relação entre a BS e a teoria de grafos, alguns conceitos matemáticos podem nos ajudar a entender e empregar esta área do conhecimento com maior domínio e propriedade. Assim, prosseguiremos com uma breve introdução sobre teoria de grafos e estrutura de rede, apresentando alguns descritores matemáticos frequentemente empregados em BS.

Uma rede (ou grafo) $G = (V, E)$ representa uma combinação de nós (V) e conectores (E) que ligam os nós. Em uma rede, o conjunto de seus nós é denotado por $V(G)$, enquanto o conjunto de seus conectores por $E(G)$. Dessa forma, o número total de nós em G é representado por n, e o número total de conectores é representado por m:

$$n(G) = |V(G)| \text{ e } m(G) = |E(G)|$$

Adicionalmente, conforme apresentado na Figura 2A-6, um conector E deve apresentar suas extremidades ligadas aos nós a e b ($a \in V$ e $b \in V$), sendo chamado eab, $E(a, b)$ ou apenas ab. Este conector pode ser representado da seguinte forma:

$$E = \{(a, b) \mid a, b \in V\}$$

As redes podem apresentar conectores diretos, ou seja, um conector orientado em determinada direção (exemplo $a \rightarrow b$, $b \rightarrow c$), sendo assim chamadas de redes direcionadas

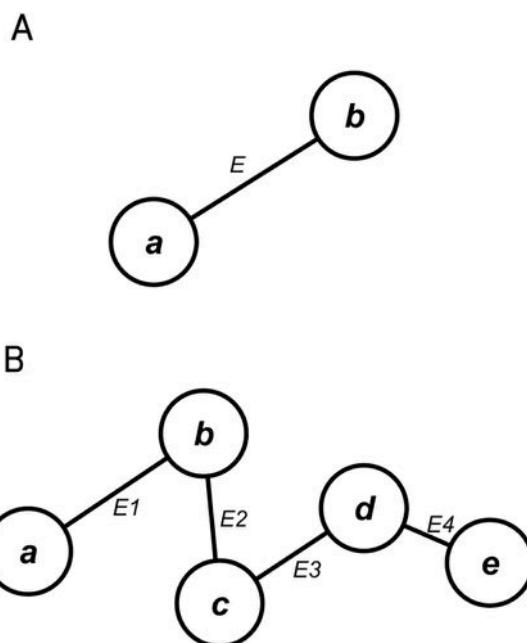
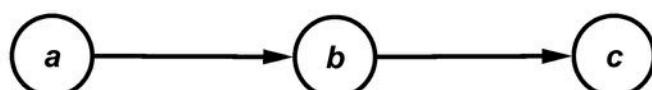



Figura 2-6: Em (A) a representação da interação de dois nós vizinhos ($V = a, b$) conectados pelo conector $E(a, b)$. Em (B) a rede pode ser descrita como $V = \{a, b, c, d, e\}$ e $E = \{ab, bc, cd, de\}$, com $n = 5$ (5 nós de a a e) e $m = 4$ (4 conectores de 1 a 4).

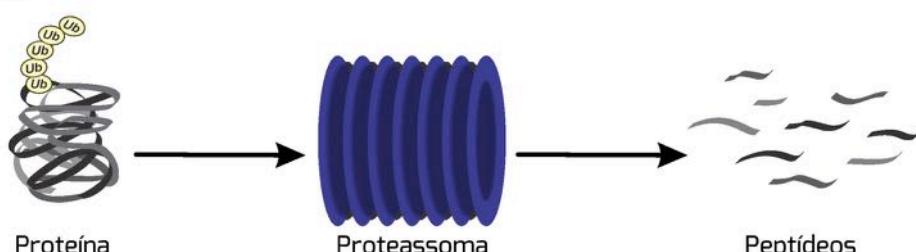
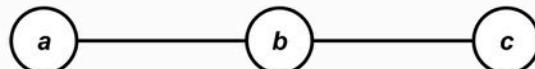



Figura 3-6: (A) Rede direta; (B) Representação da via de degradação ubiquitina-proteassoma, um dos inúmeros tipos de redes direcionadas encontradas em sistemas biológicos.

ou dígrafos (Figura 3A-6). Nos conectores E = (a, b) e E = (b, c), podemos dizer que a é antecessor a b, e b é antecessor a c. Da mesma forma, b é sucessor de a e c é sucessor de b. Um dígrafo é definido por $G = (V, E, f)$, sendo f uma função que associa cada elemento E a um par ordenado de nós em V. Uma rede representando os mecanismos de degradação ubiquitina-proteassoma de uma determinada proteína pode ser um exemplo de rede direta após o reconhecimento da proteína ubiquitinada por proteassomas, uma vez que não é possível reverter a degradação da proteína (Figura 3B-6).

Podem também existir redes não direcionadas (Figura 4A-6), que apresentam conectores orientados em ambas as direções ($a \leftrightarrow b$, $b \leftrightarrow c$), não sendo possível assim estabelecer antecessor ou sucessor. Um exemplo típico seria a reação reversível de um substrato A para um substrato B em uma via metabólica como, por exemplo, a formação de

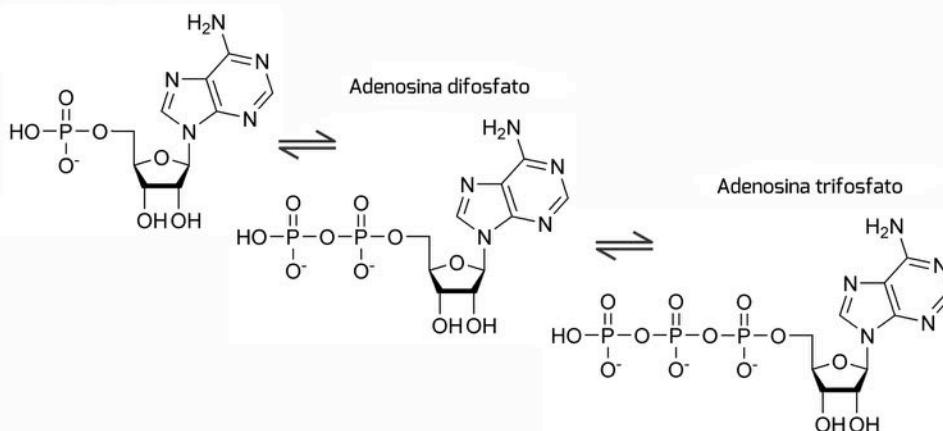



Figura 4-6: (A) Rede não direcionada; (B) Reação reversa de fosforilação e desfosforilação de adenosina difosfato, representando um exemplo de redes não direcionadas em sistemas biológicos.

diferentes moléculas fosforiladas de adenosina conforme a reação AMP↔ADP↔ATP (Figura 4B-6).

Em alguns casos, podem existir dois ou mais conectores que ligam os mesmos nós na rede. Esse tipo de interação é chamado multiconector, onde diferentes informações são representadas por cada conector, caracterizando assim um multidígrafo (Figura 5-6).

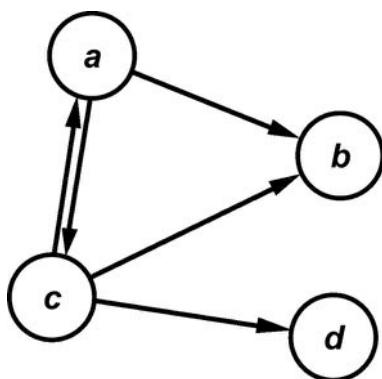



Figura 5-6: Multidígrafo $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, ac, ca, cb, cd\}$.

Observa-se, assim, que as redes apresentam interações entre os nós e que essas interações são delimitadas pelos conectores. Portanto, se $E = (a, b)$, logo os nós a e b são vizinhos ou adjacentes, e $E(a, b)$ é incidente aos nós a e b, lembrando que $E(a, b)$ se refere ao conector.

Uma das formas de representar e descrever tais interações entre os nós de uma determinada rede envolve o uso de matrizes. Assim, se considerarmos uma rede G contendo os nós v_a, ... v_n a matriz que descreve os elementos adjacentes em G é dada por:

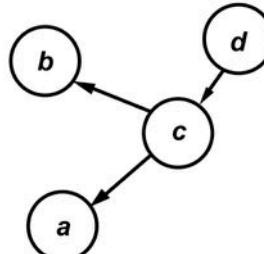
$$a_{ij} = \begin{cases} 1 \text{ se } viv_j \in E(G) \\ 0 \text{ se } v_i v_j \notin E(G) \end{cases}$$

As tabelas representadas na Figura 6-6 são um mecanismo visual para compreender como a matriz de uma rede é elaborada, tanto para redes não direcionadas (Figura 6A-6) quanto direcionadas (Figura 6B-6).

Para as redes não direcionada (Figura 6A-6) e direcionada (Figura 6B-6), as matrizes são representadas abaixo:

$$M = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \qquad M = \begin{matrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

Rede direcionada *Rede não direcionada*

Ao analisarmos uma matriz devemos considerar cada nó como uma coluna e uma linha distinta. Na análise da primeira matriz iremos interpor o nó representado na linha 1 (nó a) com o nó representado na coluna 1 (nó a) da mesma forma que as tabelas representadas na Figura 6-6, e como não há interação de a com a, nos referimos como 0. Da mesma forma, se consideramos a linha 1 (nó a) e a coluna 2 (nó b), há conexão, sendo representado por 1. Perceba que as matrizes são diferentes na rede direcionada e não direcionada devido à atribuição de uma conexão direcionada. Na matriz direcionada, tanto b está conectado a c quanto c está conectado a b. Contudo, na matriz não direcionada, somente c está conectado a b.

Também podemos definir uma rede como completa se $\boldsymbol{E}(G) = V(G)^{(2)}$, isto é, se dois nós selecionados aleatoriamente na rede G são adjacentes. Assim, uma rede completa tem n nós e é representada por K_n, sendo o número de conectores em K_n representado por $\binom{n}{2}$.

NÓS	a	b	c	d
a	0	1	0	0
b	1	0	1	0
c	0	1	0	1
d	0	0	1	0

NÓS	a	b	c	d
a	0	0	0	0
b	0	0	0	0
c	1	1	0	0
d	0	0	1	0

Figura 6-6: (A) Rede não direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, bc, cd\}$ ou $E = \{ba, cb, dc\}$, representados também na tabela pelo número 1, que indica a presença de um conector entre dois nós, exemplo $E = \{ab, ba\} = 1$. A ausência do conector entre dois nós é representada por 0. (B) Rede direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ca, cb, dc\}$. Neste caso, a tabela de interações muda devido ao direcionamento das conexões, por exemplo $E = \{ca\} = 1$, mas $E = \{ac\} = 0$.

O conjunto de nós e conectores de uma rede pode ser apresentado em uma representação mais complexa e informativa, agregando pesos (atributos) associados aos nós e conectores (Figura 7-6). Redes que apresentam nós e conectores com atributos são chamadas de redes ponderadas (G,w), onde G = (V,E) e w = V, $E \in R$, sendo R o conjunto dos números reais e w correspondente à função atributo. Por exemplo, pode-se representar uma rede neural onde o atributo indica a distância que um sinal neural deve percorrer em relação ao local de origem. Assim, se P é uma trajetória na rede, $w(P)$ é considerada a extensão de P. Redes ponderadas são amplamente usadas na bioinformática, onde G,$w(a,b)$ pode representar a quantidade e a fidelidade de informações armazenadas em bancos de dados a respeito da interação entre a e b (Figura 7-6).

Também podemos nos referir a uma rede como bipartida (Figura 8-6) onde, em G = (V, E), V pode ser dividido em V_x e V_y. Assim, cada nó de V_x é adjacente aos vértices de V_y. Desta forma, se consideramos E (a, b) significa que $a \in V_x$, enquanto que $b \in V_y$ ou $a \in V_y$ e $b \in V_x$. A aplicação de redes bipartidas na modelagem de redes biológicas pode ser vista em vários contextos, desde a análise de genótipos e SNPs (*single-nucleotide polymorphism*) em diferentes populações até a representação de conexões ecológicas e reações enzimáticas em vias metabólicas.

O modelo de redes visto até agora, na qual um conector se liga a dois nós, apesar de amplamente utilizado na avaliação da conectividade de redes biológicas, pode ser uma representação simplista quando se trata de redes metabólicas. A organização biológica que caracteriza as redes metabólicas em um contexto bioquímico consiste de complexas interações, frequentemente envolvendo diversos substratos e produtos. Para melhor representar a complexidade de reações bioquímicas, usam-se redes conhecidas como hipergrafos (Figura 9-6).

Os hipergrafos são caracterizados pela presença de hipervértices, que conectam mais de dois nós com propriedades distintas (Figu-

Figura 7-6: Representação de uma rede ponderada descrevendo: i) diferentes tipos de nós, onde cada cor representa diferentes famílias de proteínas (por exemplo, os nós verdes representam serina/treonina cinases, nós azuis representam cinases dependentes de ciclinas e nós laranjas representam as tirosina cinases); ii) diferentes tamanhos de nós, com atributo $w(a)$, representando o número de artigos w que citam a proteína a; e iii) a espessura do conector y, representando a fidelidade w da interação entre duas proteínas distintas.

Figura 8-6: Representação de uma rede bipartida, onde os nós cinzas e brancos representam diferentes grupos de uma análise. Por exemplo, cada grupo pode representar duas linhagens diferentes de *E. coli*. Para avaliar a eficiência de transformação das linhagens, estas foram divididas em quatro amostras (representadas pelos nós) e cada amostra foi incubada com diferentes plasmídeos. Os conectores apresentam os plasmídeos que obtiveram sucesso na transformação e são comuns entre as duas linhagens.

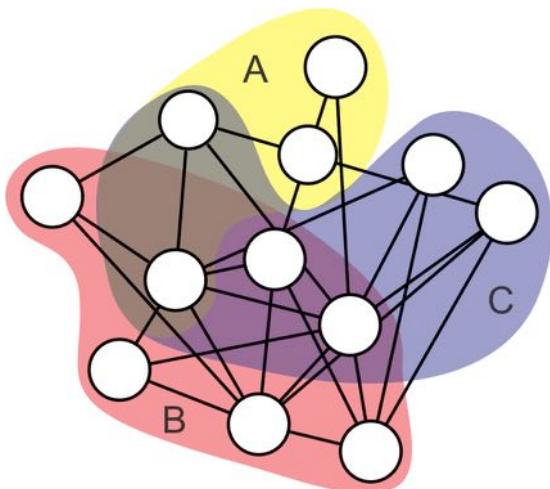



Figura 9-6: Representação de um hipergrafo. As regiões destacadas em várias cores caracterizam as diferentes propriedades ou atividades bioquímicas representadas na rede. Assim, cada cor estaria representando diferentes vias metabólicas (A, B e C). Os nós da rede indicam componentes presentes em cada uma das vias metabólicas e/ou participando de vias distintas nas regiões intersectadas.

ra 9-6). Assim, os hipergrafos são frequentemente usados em organizações bioquímicas, devido à intersecção de componentes com atividades em diferentes rotas metabólicas.

Geralmente, as redes biológicas são extensas, apresentando um grande número de nós. Contudo, análises estatísticas indicam que, dentro de uma rede maior (Figura 10A-6), podem existir redes menores que participam da composição geral e possuem maior conectividade entre si quando comparados à rede maior (Figura 10B-6). Essas subredes de $G = (V, E)$ nada mais são que uma rede $G_1 = (V_1, E_1)$, onde $V_1 \subseteq V$ e $E_1 \subseteq E$.

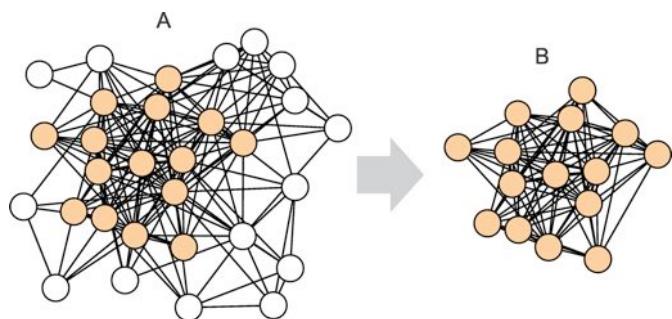



Figura 10-6: (A) Rede de interações proteína-proteína representando em laranja a subrede, o qual foi destacada em (B).

6.3. Estrutura de redes

Uma das características de uma rede é sua conectividade (também referida como grau de nó), sendo a conectividade total de uma rede definida por $C = E / N (N - 1)$, onde E representa o número de conectores e N o número total de nós.

Considere os nós V_a e V_e de uma rede. Representamos como um dos possíveis caminhos de V_a a V_e os vértices V_b, V_c e V_d, formando um conector a cada dois vértices sucessivos, caracterizados por E_1, E_8, E_4, E_7 (Figura 11-6). O nó que originou o caminho é chamado de nó inicial, enquanto que o último nó do caminho é chamado de nó final. Um caminho onde o nó inicial coincide com o nó final, sem repetições de conexões intermediárias, é chamado de circuito. Usando a mesma rede da Figura 11-6, $<d, b, c, e, d>$ formam um circuito. O comprimento de um caminho ou circuito consiste do número de conectores que pertencem ao caminho (ou circuito) ou, no caso de uma rede ponderada, pela soma dos atributos (ou pesos) dos conectores.

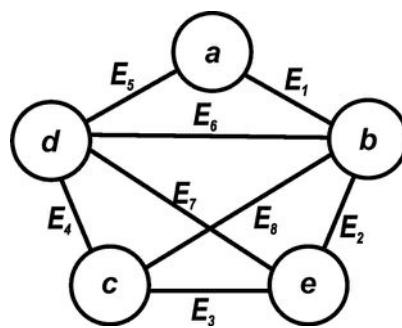



Figura 11-6: Esquema representando uma rede, onde $V = \{a, b, c, d, e\}$ e $E = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8\}$.

Um caminho de comprimento k tem exatamente $k + 1$ nós, enquanto que um circuito de comprimento k tem $k = v$ nós. Se calcularmos o comprimento de V_a a V_e, com caminho E_1, E_8, E_4, E_7 temos $k = 4$ conectores com $4 + 1$ nós. Para o circuito $<d, b, c, e, d>$ que tem como caminho E_6, E_8, E_3, E_7 temos $k = 4$ conectores, com quatro nós diferentes.

Uma importante análise em uma rede consiste em caracterizá-la conforme sua distribuição de caminhos geodésicos. Um caminho geodésico é definido como a via mais curta dentro de uma rede entre dois nós quaisquer (i e j), sendo representado por δ (i, j) em G. Um bom exemplo disso é o experimento realizado por Stanley Milgram em 1960, onde cartas foram enviadas a indivíduos aleatoriamente. A missão de cada indivíduo era enviar a sua carta a alguém que considerasse capaz de fazer com que as cartas chegassem ao seu destino final.

Essa experiência relativamente simples conclui que existem aproximadamente seis graus de separação entre dois indivíduos quaisquer no mundo. Da mesma forma, esse experimento foi a primeira demonstração significativa do efeito "mundo pequeno" (ou do inglês, *small world*), que estabelece que as redes apresentam nós conectados entre si formando um caminho mais curto entre todos os nós.

O comprimento médio de caminhos entre os nós (i, j) é definido pelo valor médio de conectores entre os nós e pode ser calculado por:

$$\delta = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{min(i,j)}$$

assumindo-se que δ_{min} (i, j) é o caminho mais curto entre os nós i e j, sendo N o número total de nós. Adicionalmente, o diâmetro da rede é definido como:

$$D = \max_{i,j} \delta_{min}(i,j)$$

e representa o maior comprimento entre dois nós. Estudos recentes têm revelado que redes biomoleculares, sociais e tecnológicas apresentam valores de comprimento médio de caminhos e diâmetro relativamente pequenos se comparados ao tamanho da rede, apresentando ordem de grandeza log (n) ou menor quando o tamanho da rede é n. Da mesma forma, a densidade de uma rede é calculada com base no número de conexões que cada nó possui, sendo definida como:

$$\rho = \frac{2m}{n(n-1)}$$

Avaliar a densidade de uma rede representa avaliar o nível de conectividade, tornando-se muito importante na definição de suas propriedades, como veremos adiante. Por exemplo, ao analisarmos a rede de interação de uma doença contagiosa, a possibilidade desta doença até então controlada tornar-se uma epidemia depende principalmente de duas variáveis: o tipo de agente infeccioso e a alta densidade de conexões (rotas de transmissão). O procedimento de quarentena (isolamento) quando um determinado indivíduo apresenta os sintomas da doença é justamente reduzir a conectividade da rede de transmissão.

Alguns modelos de rede (como as redes de livre escala e hierárquica, discutidas adiante no item 6.5.) podem apresentar clusterização, isto é, os nós tendem a se agrupar. Isso significa que se um nó A se liga ao nó B, e o nó B se liga ao nó C, então há grandes chances de A se ligar a C também. Assim, a rede é composta de centenas de triângulos, ou seja, grupos de três nós conectados entre si, onde cada lateral de um triângulo pode pertencer a outro triângulo.

Podemos quantificar a fração de triplos nós que apresentam um terceiro conector preenchendo um triângulo pelo coeficiente de clusterização:

$$C = \frac{3 \times \text{número de triângulos na rede}}{\text{número de nós triplamente conectados}}$$

Na equação, o número três presente no numerador é devido ao fato que cada lateral de um triângulo contribui com outros três triplos nós, além de garantir que C seja $0 \leq C \leq 1$. Dessa forma, o coeficiente de clusterização avalia a probabilidade dos nós i e j serem vizinhos, já que ambos são vizinhos do nó h. Assim, o coeficiente de clusterização local de um nó i pode ser determinado por:

$$C_i = \frac{2e}{k(k-1)}$$

onde um nó i tem k vizinhos com e conexões entre eles. Contudo, pode-se também atribuir o coeficiente de clusterização média para a rede total, sendo definido por:

$$C = \frac{1}{N} \sum_i C_i$$

Ao analisarmos uma rede de processos biológicos, notamos que esta apresenta um maior coeficiente de clusterização média quando comparado a uma rede aleatória. Isso possivelmente se deve ao fato de pro-

cessos celulares ocorrerem de forma dependente da organização de diversos subconjuntos (*clusters*) de biomoléculas.

Em uma rede consideramos como sendo o grau de um nó o número de conectores k que incidem a este nó. Assim, a distribuição do grau $P(k)$ é definida por ser uma fração de nós com grau k dentro de uma rede. Então sendo $k = 0, 1, 2,...$ $P(k)$ indica a probabilidade de determinado nó ter grau k. A distribuição de grau é definida por:

$$P(k) = \frac{n_k}{n}$$

onde temos n nós na totalidade da rede e n_k representa a quantidade de nós com grau k.

Uma rede aleatória que apresenta n nós conectados ou não com probabilidade p, tem uma distribuição binominal de grau com parâmetros $N - 1$ e p:

$$P(k_i = k) = C_{N-1}^{k}\, p^k (1-p)^{N-1-K}$$

Outras redes, no entanto, tem distribuição de grau bem diferente. Redes de livre escala (como a maioria das redes biológicas) apresentam distribuição do grau que segue uma Lei de Potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$ (ver adiante).

Outra estimativa numérica pode ser feita, a função de distribuição cumulativa avalia a probabilidade de um nó ter um grau maior do que k:

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

Agora, o que aconteceria se, por acaso, resolvessemos excluir alguns poucos nós da rede? Certamente iríamos alterar o comprimento de alguns caminhos e circuitos da rede de forma pouco significativa. Contudo, se formos excluindo mais nós, progressivamente, veremos que a comunicação da rede fica cada vez mais esparsa, até se tornar desconectada. A capacidade de uma rede de tolerar a deleção de nós é chamada de resiliência.

Em 2000, um estudo conduzido por Albert-László Barabási e colaboradores mostrou que a Internet pode ser altamente resiliente na remoção de nós aleatórios. Isso se deve ao fato de que a quantidade de nós com baixo grau de interação é maior em uma rede do que nós com alto grau de interação. Em compensação, se a remoção iniciar a partir dos nós com mais alto grau de interação, a alteração será brusca. Neste caso, observa-se um aumento da distância entre os nós, de forma que apenas poucos nós precisam ser removidos para destruir a comunicação da rede. Assim, fica claro que a Internet apresenta baixa resiliência na remoção de nós com alto grau, tornando-se vulnerável a ataques de *hackers*.

Outro exemplo seriam as redes de interação proteína-proteína. Estas redes geralmente apresentam muitas proteínas com poucas interações e algumas proteínas possuindo muitas interações (chamadas de *hubs*, ver adiante). Desta forma, redes de interação proteína-proteína são resilientes à deleção de nós aleatórios, porém extremamente vulneráveis a ataques em proteínas *hubs*.

Os nós de uma determinada rede podem apresentar tendências de conexão. Em outras palavras, duas redes completamente diferentes topologicamente podem apresentar a mesma distribuição do grau. Assim, em uma rede é preciso considerar o padrão de correlação do grau dos nós, onde a conectividade de um nó reflete nas suas possibilidades de ligação.

A tendência de conexão que uma rede apresenta pode ser chamada de assortatividade e desassortatividade. A assortatividade significa que os nós de uma rede apresentam uma tendência a interagirem com outros nós semelhantes, por exemplo, nós do tipo A interagem preferencialmente com nós também do tipo A (Figura 12A-6). Vértices com alto grau tendem a interagir com vértices que também apresentam alto grau. No entanto, chamamos de desassortatividade se os nós de uma rede interagem preferencialmente com nós diferentes dele mesmo, por exemplo, nós do tipo A tendem a interagir com nós do tipo B. Neste caso, um nó com alto grau tem tendência a interagir com nós que apresentem baixo grau (Figura 12B-6).

A correlação de grau dos nós i e j é feita por distribuição de probabilidade conjunta $P(k_i, k_j) = P(k_i)\, P(k_j)$. Podemos ainda calcular a assortatividade ou desassortatividade da rede como um todo, considerando:

$$r = \frac{\sum_i eii - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

Se $r = 1$ a rede é considerada assortativa, enquanto que se $r = -1$, a rede é completamente desassortativa.

Caracteristicamente, redes assortativas são mais resilientes e apresentam *hubs* bem conectados, enquanto que redes desassortativas são redes mais vulneráveis com nós conexos a *hubs* esparsos (Figura 12-6).

A conectividade de uma rede também pode ser avaliada pela teoria da percolação. Essa teoria tem por objetivo estudar a conectividade da rede pela avaliação de sua arquitetura, caracterizando a distribuição do tamanho dos *clusters* e descrevendo como ocorre a transferência de informações, por exemplo, de A para B.

Redes aleatórias caracteristicamente apresentam baixa tendência em possuir pequenos *clusters* isolados e uma grande probabilidade em formar um componente conectado gigante. Como visto anteriormente, determinadas redes são altamente resilentes à deleção aleatória de nós. A variação na fração dos nós no maior componente da rede (componente gigante) é a forma mais fácil de calcular a resiliência. Imagine dois nós conectados na rede. Se estes nós pertencem a um componente gigante, há grande probabilidade de se comunicarem com uma extensa proporção de nós da rede. No entanto, nós que participam de pequenos componentes comunicam-se apenas com uma parte reduzida da rede. Essa capacidade de comunicação é responsável pela forma como a informação é transferida de um ponto a outro. Assim, associamos a resiliência com a percolação local (refere-se aos nós), enquanto que a percolação de ligação (refere-se aos conectores) está relacionada ao processo de dispersão (Figura 13A-6).

Também podemos considerar os nós de uma rede como ocupados (funcionais) ou desocupados (falhos), dependendo da sua funcionalidade. A probabilidade de um nó estar ou não ocupado pode ser uniforme ou pode depender do grau do nó, sendo que os nós funcionais da rede formam o componente gigante em um modelo de percolação. Assim, os nós ou conectores falhos não participam da transferência de informação, e igualmente, não participam do componente gigante (Figura 13B-6). Dessa forma, ao observar a propri-

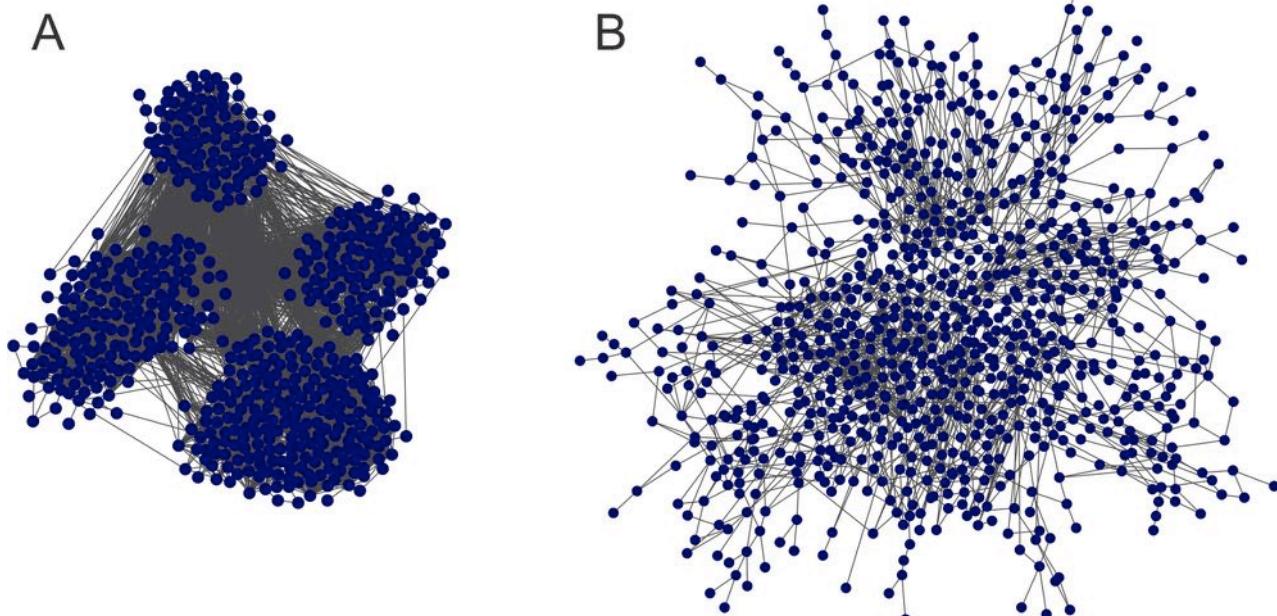



Figura 12-6: Ilustração representando em (A) uma rede assortativa com nós bem conectados que apresentam conexões com outros nós também fortemente conectados. Em (B), uma rede desassortativa, onde os poucos nós que apresentam mais conexões interagem com nós menos conectados, resultando em uma rede menos densa.

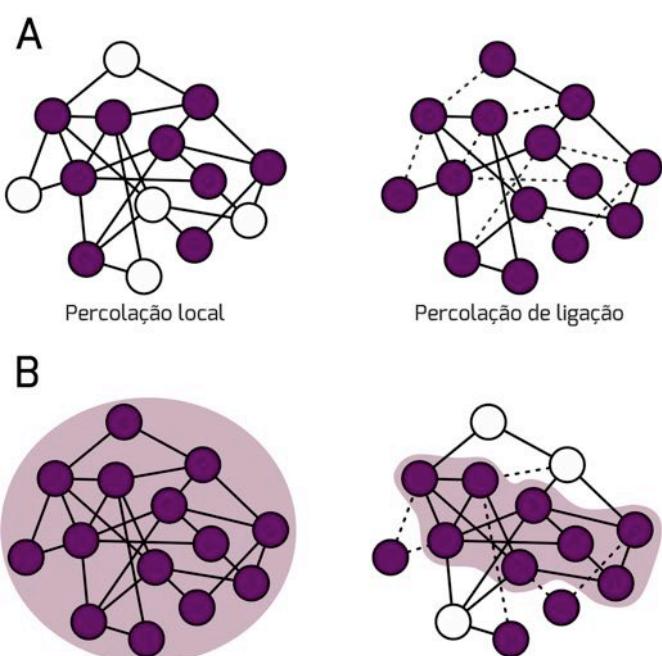



Figura 13-6: (A) Redes de percolação local e de ligação, onde os nós sólidos estão ocupados ou funcionais, enquanto que os nós brancos são desocupados ou falhos. (B) Representação do componente gigante. Após o surgimento de nós e conectores falhos, sua proporção é alterada e, por conseguinte, as possibilidades de transferência de informações.

edade de percolação de um *cluster*, considerando uma probabilidade de ocupação variável, podemos determinar que isso afeta diretamente a conectividade de uma rede, tornando-a altamente resiliente ou não. Porém, ao combinarmos a percolação local e de ligação, teremos um modelo robusto contra falhas de nós ou conectores.

Os modelos de percolação são utilizados em muitas redes, porém um dos modelos mais interessante é o da dispersão de uma doença. Nesse modelo, cada nó representa o hospedeiro e os conectores representam a capacidade de transmissão da doença entre um hospedeiro e outro. O nó (indivíduo hospedeiro) está ocupado se for suscetível à doença, enquanto que um nó que representa um indivíduo que tomou a vacina seria considerado como desocupado. Da mesma forma, os conectores são considerados ocupados se há possibilidade de transmissão (Figura 14-6). Levando em conta este modelo, o início de uma epidemia representa a transição de percolação.

Apesar de ter sido originalmente desenvolvida com o objetivo de responder às perguntas em química orgânica, os modelos de percolação têm sido usados com sucesso para estudar diversos fenômenos, como transferência de sinal em neurônios e condutividade elétrica. Em 1987, Robert H. Gardner foi um dos primeiros pesquisadores a usar a teoria de percolação na Ecologia da Paisagem, sendo útil também na avaliação de corredores ecológicos e redes de incêndios florestais.

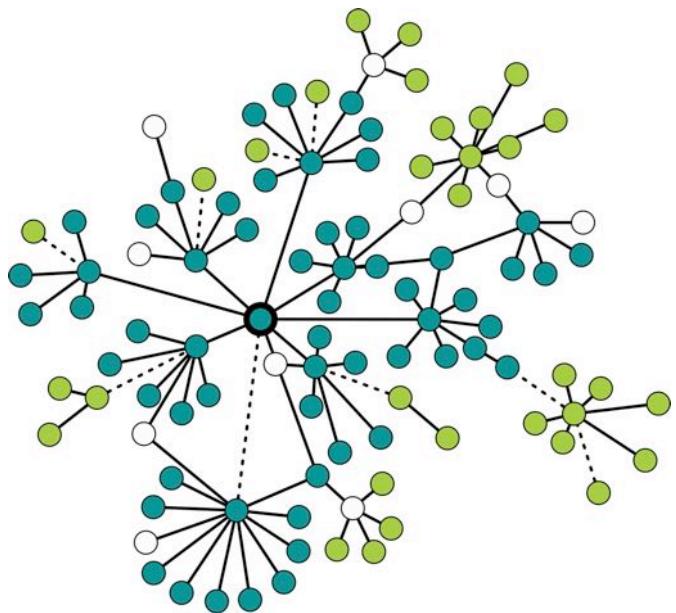



Figura 14-6: Modelo simplificado de dispersão de uma doença considerando um grupo de trabalho em uma empresa. Suponhamos que o indivíduo central contraiu uma doença viral de fácil transmissão, como a gripe simples. Assim, todos os indivíduos com os quais ele entrou em contato neste período também contraíram a doença (nós azuis), com exceção daqueles que foram vacinados (nós brancos). Neste caso, além de não contraírem a doença, também não a dispersaram. Os conectores pontilhados indicam que não houve interação física durante o período passível de contrair a doença entre o indivíduo saudável com o contaminado. Desta maneira, os indivíduos representados pelo nó verde claro, apesar de não terem sido vacinados, não contraíram a doença por não entrarem em contato com indivíduos contaminados.

6.4. Propriedades de rede

Diversas propriedades são regularmente empregadas na análise de redes biológicas, cada uma fornecendo informação sobre as interações e/ou componentes de um determinado sistema. Estas propriedades podem ser referentes a nós individuais, isto é, grau de nó ou *node degree*, ou podem contemplar a rede como um todo como é, por exemplo, o caso da modularização e do diâmetro da rede.

Em uma análise de biologia de sistemas, a análise estatística destas propriedades possui papel crítico na geração de dados conclusivos e confiáveis, constituindo-se assim em redes capazes de descrever com alto grau de fidelidade um determinado modelo biológico, de identificar alvos proteicos críticos na rede ou no desenvolvimento de caminhos moleculares.

Modularidade

Uma das principais características quando nos referimos a propriedades da topologia de redes é a chamada modularidade ou clusterização. O conceito de modularidade é antigo e já amplamente usado em outras áreas do conhecimento, como nas ciências sociais. Dentro das ciências biológicas, é um conceito comum nas áreas da biologia evolutiva, biologia molecular, biologia de sistemas e biologia do desenvolvimento.

Todas as ideias de modularidade giram em torno do conceito de padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados. De forma geral, a modularidade é um princípio de união entre diferentes tipos de elementos e conexões naturalmente formadas no meio biológico, como na interação entre indivíduos de mesma espécie. Um exemplo é a *Pollenia rudis*, uma espécie de mosca conhecida como *cluster fly* em decorrência de seu hábito de se agrupar com indivíduos da mesma espécie.

Este princípio é visto em todos os lugares, seja na nossa tendência de formar sociedades e grupos preferenciais de interação interpessoais ou na nossa tendência de organizar objetos por seu tipo, função e cores, dentre outros. Em nível molecular é visto, por exemplo, em elementos que atuam num mesmo processo biológico, como conjuntos de moléculas de RNA responsáveis pela degradação e síntese de ácidos nucleicos ou grupos de proteínas que atuam num mesmo processo biológico como a replicação de DNA e a transcrição gênica.

Existem dois tipos distintos de módulos:

i) Módulo Variacional: apresenta características que variam entre seus componentes e são relativamente independentes de outros módulos, porém possuem um número considerável de ligações com outros módulos;

ii) Módulo Funcional: possui elementos que normalmente atuam juntos em alguma função fisiológica distinta e são semiautônomos (*quasi-autonomous*) de outros módulos. Esses módulos compreendem a maioria dos módulos vistos em redes biológicas.

Módulos variacionais podem ser exemplificados na Figura 15B-6 e C, representando a formação de uma mandíbula de rato. Apesar de se tratar da diferenciação de um tecido, podemos usá-la como modelo variacional devido ao fato de diferentes proteínas e genes serem responsáveis pela formação de uma unidade estrutural única (o ramo ascendente e da região alveolar). Desta maneira, é uma unidade estrutural (um único osso) que se origina de diferentes módulos. Assim, o módulo variacional consiste numa integração de vários de genes que dividem efeitos pleiotrópicos entre si e que possuem poucos efeitos pleiotrópicos com outros *clusters*, sendo praticamente independente.

Módulos de genes de desenvolvimento embrionário, relacionados à diferenciação ou formação de padrões corporais, tendem a ser quase independentes de outros módulos, uma vez que erros na sua expressão ou atuação podem ser letais para o embrião. Por isso, esses módulos de desenvolvimento tendem a depender de elementos dentro do próprio

grupo para sua expressão. Podemos visualizar um exemplo de um módulo funcional na Figura 15A-6.

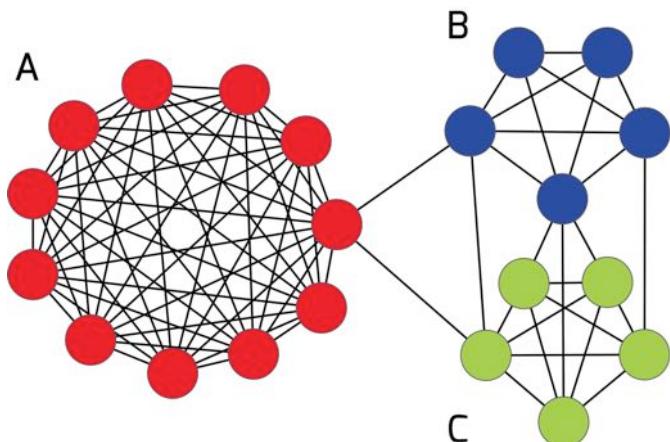



Figura 15-6: Exemplos de uma rede com diferentes módulos representados. Os módulos variacionais B (azul) e C (verde) se encontram praticamente independentes do módulo A (vermelho), porém possuem proteínas em comuns entre si. Contudo, o módulo A pode ser considerado funcional, uma vez que possui apenas uma conexão com cada outro módulo, sendo praticamente independente.

Ao determinarmos a quantidade e o tipo de módulos presentes em uma rede devemos levar em consideração o coeficiente de agrupamento (C_i) ou clusterização. O coeficiente analisa a tendência de um nó de se associar com seus vizinhos ("*cliquishness*"), onde "clique" é definido como um grafo maximamente conectado.

Como mencionado anteriormente, a clusterização é dada pela fórmula $C_i=2n/k_i(k_i-1)$, onde k_i é o tamanho da vizinhança de vértices (nós) do vértice i, e n é o número de conectores na vizinhança. Assim, quanto maior o coeficiente de clusterização, mais conectado é o *cluster*. Evolutivamente, as proteínas que compõem módulos altamente agrupados tendem a ser conservadas ou perdidas juntamente, caso haja uma variação dentro do grupo.

Outro conceito essencial para entender a formação de um *cluster* em um sistema biológico é a presença de *hubs*. Os *hubs* podem ser classificados em dois grupos:

i) *party hubs*, proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligadas no mesmo tempo e/ou espaço,

ii) *date hubs*, que são *hubs* que se ligam a diferentes proteínas em diferentes módulos (inter-módulo), ou seja, diferentes tempo e/ou espaços, consequentemente apresentando um papel global na rede (Figura 16-6). Estes termos podem ainda receber denominações específicas no contexto do conceito de centralidades (ver adiante).

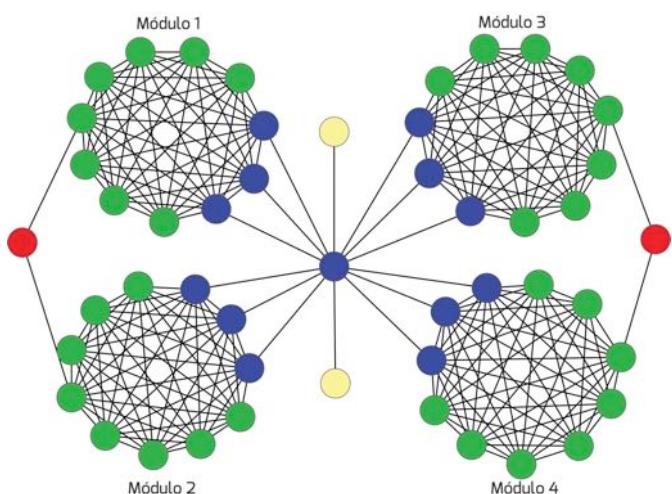



Figura 16-6: Diferentes tipos de centralidade em uma rede biológica. Em verde são apresentadas proteínas envolvidas em *party hubs* e encontradas em módulos. Em amarelo encontram-se as proteínas não-*hub*/não-gargalo, que são aquelas que não possuem alto valor de grau de nó ou *betweenness*, sendo consideradas componentes funcionais dos módulos. Em azul estão as proteínas *hub*-gargalo (*date-hub*) que possuem alto valor de grau de nó e de *betweenness*, sendo consideradas fundamentais para o funcionamento de redes. Em vermelho estão identificadas as proteínas do tipo gargalo, com alto valor de *betweenness* e essenciais na ligação entre módulos e processos biológicos.

Os *party hubs* são componentes clássicos de módulos funcionais, uma vez que estes são quase independentes de outros módulos, enquanto *date hubs* são fundamentais para módulos variacionais, pois estes se ligam a

outros módulos.

Assim, uma mutação em um *party hub* vai afetar principalmente as proteínas referentes ao seu próprio módulo, enquanto a mutação em um *date hub* (Figura 16-6) pode afetar vários módulos. Contudo, não existe diferença de importância entre *party* ou *date hub*. A deleção de um *hub* em um módulo funcional pode ser tão letal quanto a deleção em um módulo variacional.

Baseado em dados estruturais, os *hubs* podem ser ainda classificados em *singlish* (com uma ou duas interfaces) e multi-interface (com mais de duas interfaces). *Hubs* com interface *singlish* somente se ligam a outras proteínas de maneira alternada e transitória, enquanto *hubs* multi-interface se ligam a diferentes proteínas concomitantemente.

Ontologias Gênicas

Nos últimos anos, o desenvolvimento e uso de técnicas de análise como microarranjos, ChIP-chip e espectrometria de massas e suas aplicações no estudo de cada vez mais organismos gerou um grande acúmulo de dados genômicos e proteômicos. A leitura e interpretação simples e concisa destes vem requerendo o desenvolvimento de novas abordagens, contexto no qual, em 1990, foi criado o chamado *Gene Ontology Project*.

Ontologia gênica refere-se ao produto de um determinado gene e à função que ele desempenha na maquinaria celular. São classificadas em três níveis hierárquicos:

i) Componente celular, descrevendo a localização da proteína na célula;

ii) Processo biológico, referindo-se à série de eventos realizados por uma ou mais funções celulares;

iii) Função molecular, descrevendo a atividade que uma dada proteína desempenha no meio celular.

Essas informações são guardadas em forma de "anotações ontológicas", onde cada uma possui um número de identificação e se encontram disponíveis em bancos de dados como www.geneontology.org.

Da mesma forma, essas anotações não são restritas a humanos, mas abrangem diversos organismos modelo como *Mus musculus*, *Gallus gallus*, *Saccharomyces cerevisiae*, *Caenohabditis elegans* e *Escherichia coli*, além de outros organismos não-modelo mas que já possuem alguma anotação.

De um modo geral, a ontologia gênica tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um mesmo processo biológico. Em biologia de sistemas o emprego de ontologias gênicas pode se mostrar muito útil para direcionar a análise da rede, possibilitando a verificação dos tipos de processos biológicos existentes na rede e das proteínas presentes. Um modelo hipotético de como uma rede poderia se apresentar em termos de ontologias gênicas se encontra na Figura 17-6, onde diferentes nós poderiam estar relacionados a diversos processos.

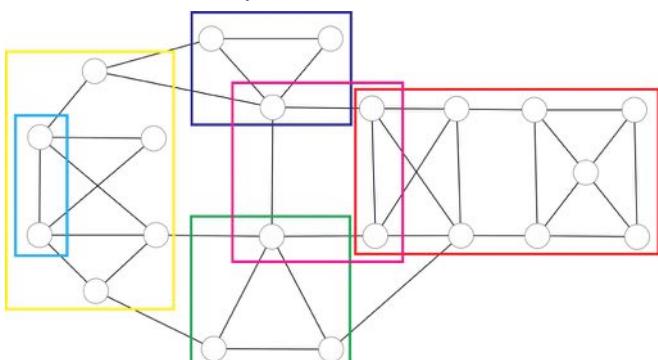



Figura 17-6: Modelo hipotético da presença de ontologias gênicas em uma rede. Na figura acima, cada cor representa um processo identificado. É importante ressaltar que uma proteína pode estar presente em mais de uma ontologia. Da mesma forma, uma ontologia pode estar dentro de outra. Como por exemplo, o quadrado amarelo poderia significar transcrição, enquanto o quadrado azul claro (inserido no amarelo) poderia significar apenas o complexo de iniciação da RNA polimerase II.

A Figura 18-6 mostra um exemplo de aplicação de ontologias gênicas em uma rede biológica. Nessa análise foi utilizado o programa *Biological Network Gene Ontology*

(BiNGO) 2.44, um *plug-in* do programa Cytoscape. É possível, assim, identificar proteínas ou genes com efeitos pleiotrópicos, a saber: a proteína Tp53, a proteína *breast cancer 1* (BRCA1) e a proteína *bloom syndrome protein* (BLM), as quais se encontram nas três ontologias da rede (reparo de DNA, regulação positiva da transcrição e ciclo celular).

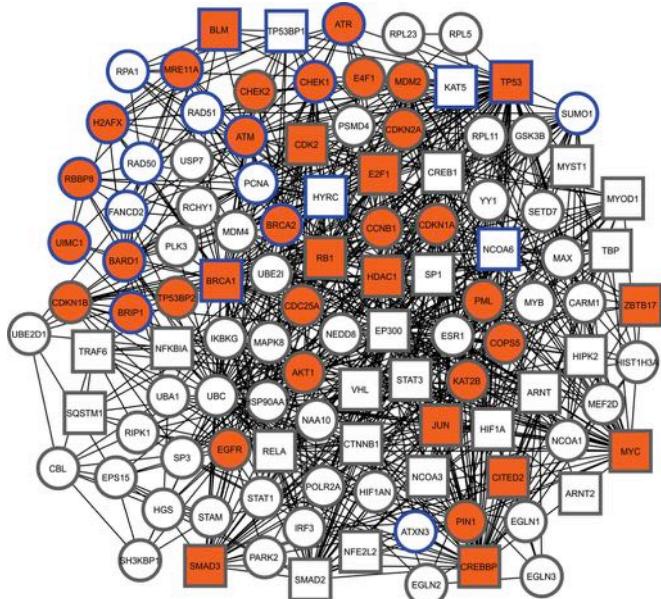



Figura 18-6: Exemplo de uma rede analisada pelo *plugin* BiNGO 2.44, o qual analisa as principais ontologias gênicas. A rede mostra três processos biológicos (GOs): *i*) Regulação do ciclo celular (nós de cor laranja); *ii*) Regulação positiva da transcrição (nós de formato quadrado); *iii*) Resposta a dano de DNA (nós com a linha azul). É possível observar que mais de um nó compõe diferentes GOs.

Centralidades para nós

Como vimos até então, a grande vantagem da biologia de sistemas é permitir a visualização dos componentes moleculares de um sistema biológico de forma dinâmica e global. Contudo, quando falamos de uma rede, temos que levar em consideração todas suas estruturas, como *hubs* e módulos. Deste modo, o objetivo da análise de centralidades é procurar o elementos mais importantes na topologia geral da rede.

Grau de nó

Um dos parâmetros básicos de análise topológica é o parâmetro de grau de nó (ou *node degree*), referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó. Esses nós que apresentam uma grande quantidade de conexões são chamados de *hubs*, os quais são conectados a outros *hubs* ou nós com menos conexões (Figura 16-6). Como veremos posteriormente, uma rede de livre escala é definida por uma lei de potenciação, o que significa que essa rede terá poucos nós altamente conectados. O grau de nó é referente ao valor distribuição de nó, $P(k)$, que informa a probabilidade de um nó ter k conexões, conforme visto em *Estrutura de redes.*

Numa visão biológica, podemos exemplificar um *hub* como uma proteína que se liga a várias outras e acaba possuindo uma função regulatória importante na rede. Normalmente, proteínas consideradas apenas *hubs* se encontram dentro de módulos. A perda de conexões de uma proteína *hub* pode lhe tirar esta condição modular. Sua deleção em uma rede de interação proteína-proteína poderia afetar a ação de diversas proteínas vizinhas e até mesmo na formação de módulos.

Betweenness

O parâmetro denominado *betweenness* é definido como o número de caminhos mais curtos que passam por um único nó, estimando a relação entre eles. Por exemplo, para calcular o valor de *betweenness* de um nó n é calculado o número de caminhos mais curtos entre i e j, e a fração deste caminhos que passam pelo nó n. Deste modo, um nó n pode ser atravessado por diversos caminhos alternativos, que ligam i e j.

Matematicamente, o valor de *betweenness* é dado pela seguinte fórmula:

$$Bet(n) = \sum_{i \neq n \neq j \in V} \frac{\sigma_{ij}(n)}{\sigma_{ij}}$$

onde σ_{ij} representam caminhos geodésicos entre os nós i e j, e $\sigma_{ij}(n)$ é o total destes caminhos mais curtos

que passam por n.

Por exemplo, uma proteína com alto valor de *betweenness* apresentaria uma elevada capacidade de interação e/ou sinalização com outras proteínas, processos biológicos ou *clusters*. Uma proteína com tais características é chamada de *bottleneck* ou gargalo. Na Figura 16-6, temos dois exemplos de uma proteína com alto valor de *betweenness*.

Não existe uma maneira óbvia de se encontrar proteínas gargalo. Porém, é possível que rotas de sinalização possuam grande incidência de proteínas gargalo, uma vez que são necessárias para sinalização entre compartimentos e processos biológicos distintos. Contudo, proteínas gargalo não necessariamente possuem um grande número de interações com outras proteínas.

Closeness

O valor de *closeness* pode ser entendido como o caminho mais curto entre um nó n e todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó (Figura 19-6). Um alto valor de *closeness* indica que todos os outros nós estão próximos do nó n, enquanto que um baixo valor indicaria que os outros nós encontram-se distantes.

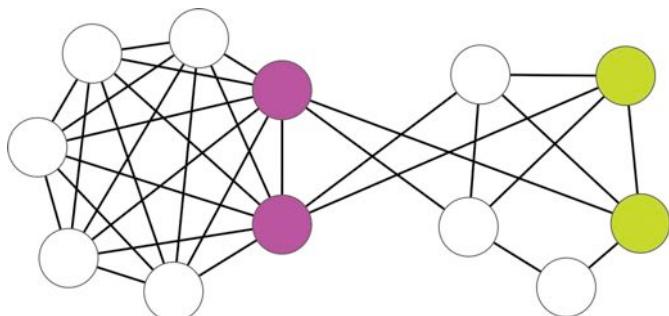



Figura 19-6: Caracterização de nós com diferentes valores hipotéticos de *closeness*. Os nós em roxo, dadas as suas maiores conectividades com a rede no geral, possuem um valor maior de *closeness*, enquanto que os nós em verde, por possuírem poucas conexões com a rede, apresentam baixo valor de *closeness*.

Este parâmetro é dado pela fórmula:

$$Clo(v) = \frac{1}{\sum w \in v^{dist\,(v'w)}}$$

onde o valor de *closeness* de um nó v $[Clo(v)]$ é determinado através do cálculo e somatório dos caminhos mais curtos entre um nó v e todos outros nós w $[dist(v,w)]$ dentro da rede.

Uma proteína com alto valor de *closeness* poderia ser considerada relevante para muitas proteínas, porém irrelevante para outras. Em termos biológicos, ela seria importante na regulação de muitas proteínas, porém sua atividade pode não influenciar outras. Ao compararmos essas informações com módulos podemos dizer que uma rede com uma média de *closeness* alta é mais provável de estar organizada como um módulo funcional, enquanto uma com baixo valor de *closeness* é mais provável de estar organizada como um módulo variacional.

Diâmetro

O diâmetro pode ser considerado um dos primeiros parâmetros referentes à “compactação”, isto é, proximidade dos nós da rede. Ele indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo. Deste modo, uma rede com baixo diâmetro é considerada mais completa, uma vez que suas proteínas estão mais interligadas entre si.

Um baixo diâmetro pode indicar que as proteínas de uma determinada rede possuem uma maior facilidade de se comunicar e/ou influenciar umas as outras, apontando para uma relação funcional co-evolutiva (Figura 20-6).

Os parâmetros de centralidades podem ser alterados com a adição ou deleção de nós ou conexões na rede (Figura 21-6). Como já mencionado, em um sistema molecular, a perda de uma conexão pode ser considerada a mudança de um domínio, impedindo a ligação

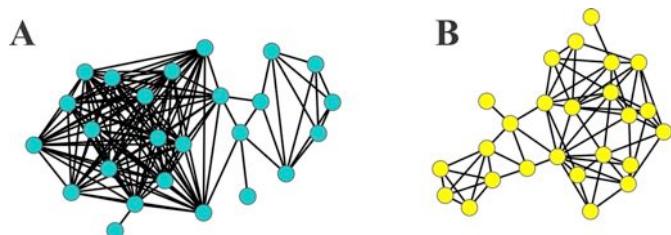



Figura 20-6: Em (A) uma rede com alto diâmetro e em (B) rede com baixo diâmetro. Pelo fato dos nós da figura A estarem mais interligados entre si, a rede é considerada mais "compacta", pois seus nós mais facilmente podem influenciar uns aos outros. Entretanto, em B, a rede possui muito menos conexões, portanto a deleção de um nó irá afetar a rede de um modo mais sutil.

de duas proteínas ou a mudança de um produto gênico, criando proteínas anormais que não mais farão as mesmas conexões. Contudo, mudanças topológicas nas redes biológicas são processos normais durante a evolução. A deleção e a duplicação de um gene, assim como a perda de interações, sejam pela mudança estrutural ou de função, são processos muitas vezes selecionados e necessários para sobrevivência celular.

Centralidade para conectores

Os elementos mais informativos de uma rede de interação podem ser avaliados através da análise da centralidade. Dentre as possíveis centralidades avaliadas, o *betweenness* de um conector pode medir a influência de certos conectores no fluxo de informações entre os componentes da rede.

O *betweenness* de um conector e é simplesmente o número de caminhos mais curtos entre pares de nós que percorrem e. Se uma rede contém módulos que são conectados por poucos conectores intermodulares, então os caminhos mais curtos entre os diferentes módulos devem passar por estes poucos conectores. Assim, os conectores unindo módulos terão altos valores de *edgebetweenness* (Figura 22-6).

Neste caso, os pares de nós unidos pelos conectores serão de diferentes módulos. Se o valor de *edgebetweenness* de um conector é baixo, esse conector provavelmente fará parte do módulo, uma vez que dentro do módulo os nós são mais interligados entre si. Portanto, *edgebetweenness* é a frequência de um conector que se coloca sobre os caminhos mais curtos entre todos os pares de nós. Em

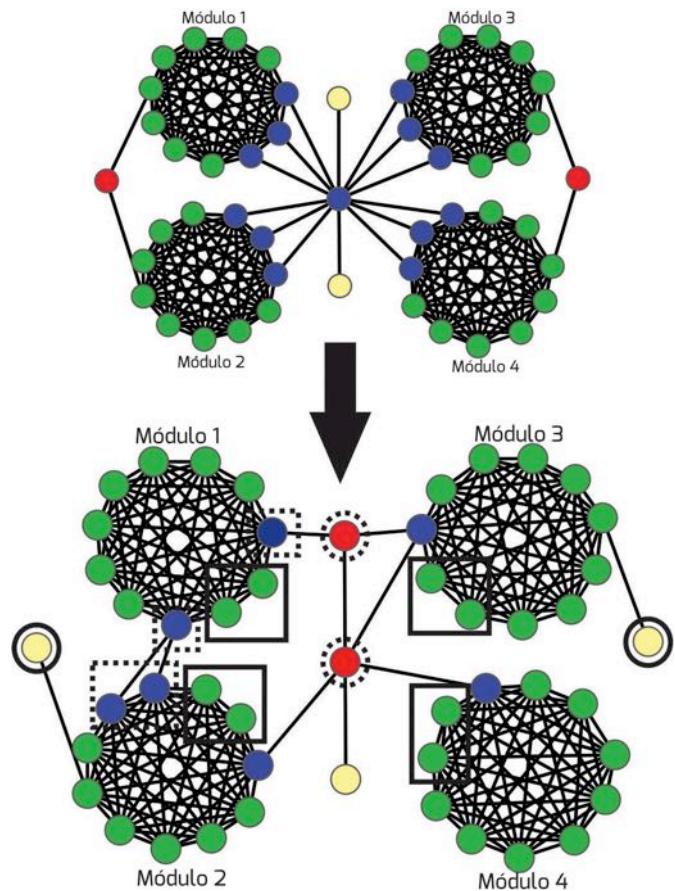



Figura 21-6: Modificações na topologia de rede podem alterar as centralidades. Devido à perda de conexões com nós fora do módulo, os nós marcados pelos quadrados foram transformados em *party-hubs* (nós verdes), deixando de ser *hubs*-gargalos (nós azuis). Porém, marcados pelos quadrados pontilhados, há nós que além de ganharem conexões, passaram a se ligar a outros módulos, saindo do estado de não-*hub*/não-gargalo para *hub*-gargalo (nós amarelos). Marcados por círculos, os nós antes gargalos (nós vermelhos), agora pela perda de uma conexão, se tornam não-*hubs*/não-gargalos. Por fim, os nós marcados pelos círculos pontilhados, devido à perda de muitas conexões (nó central) e ao ganho de uma conexão (nó acima), se tornam gargalos, perdendo os status de *hub*-gargalo e de não-*hub*/não-gargalo respectivamente.

Figura 22-6: Representação de *edgebetweenness*. Conectores em vermelho apresentam valores altos de *betweenness*, pois representam o caminho mais curto do fluxo de informação entre os três módulos representados.

uma rede proteica, um conector com alto valor de *betweenness* provavelmente representa o caminho mais curto de comunicação entre dois processos biológicos.

Como conectores com altos valores de *betweenness* são mais prováveis por posicionarem-se entre módulos, a remoção sucessiva destes conectores pode eventualmente isolar estes mesmos módulos. Essa desordem na rede, conforme será visto adiante, é conhecida como perturbação de conector.

6.5. Tipos de redes

Rede Aleatória

Os matemáticos Paul Erdös e Alfréd Rényi iniciaram seus estudos sobre redes aleatórias em 1960. Este modelo de rede tem impulsionado o interesse de diversos cientistas ao longo dos anos por ser um dos primeiros modelos de rede descoberto. Porém, apesar de amplamente estudadas, redes aleatórias não capturam a realidade de um sistema biológico (Figura 23-6).

Essas redes consistem de N nós, com cada par de nós conectados (ou não) com

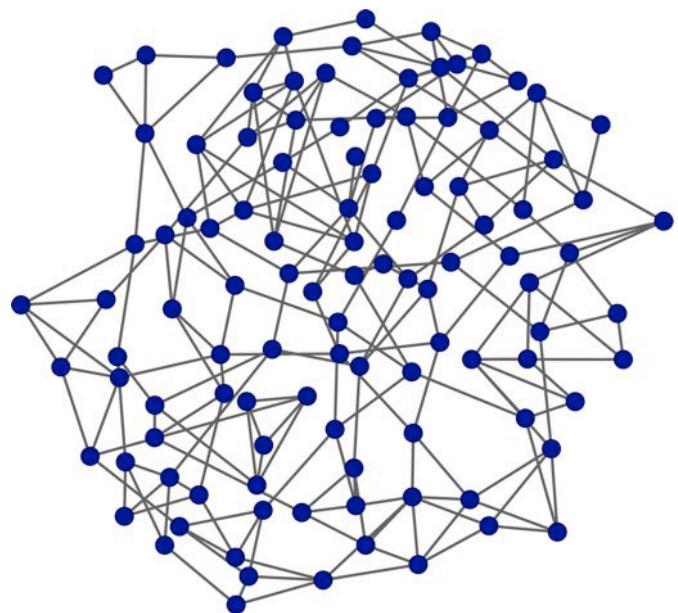



Figura 23-6: Ilustração de uma rede aleatória consistindo em 109 proteínas. A rede apresenta $P(k)$ 3,8. Observe que as conexões de cada nó são valores próximos a 4, o que está de acordo com $k \approx <k>$.

probabilidade p, gerando uma rede de conexões aleatórias com aproximadamente $pN\,.\,(N - 1) / 2$. Dessa forma, o grau dos nós segue uma distribuição de Poisson com máxima em $<k>$ e a maioria dos nós apresentando aproximadamente o mesmo número de conexões $k \approx <k>$, com grau próximo ao da média da rede. Raramente surgem nós que apresentam mais ou menos conexões que $<k>$. Adicionalmente, redes aleatórias apresentam a propriedade "mundo pequeno" e distribuição de grau exponencial, sendo estatisticamente homogêneas.

Rede de livre escala

O modelo de rede de livre escala foi introduzido por Barabási e Albert em 1999 onde se observa que redes complexas, como as redes de citações de artigos científicos, redes metabólicas, redes socais e a World Wide Web apresentam distribuição de grau que segue uma lei de potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$. Essas redes são consideradas como livres de escala (Figura 24-6) pois a lei de potência não permite uma escala característica.

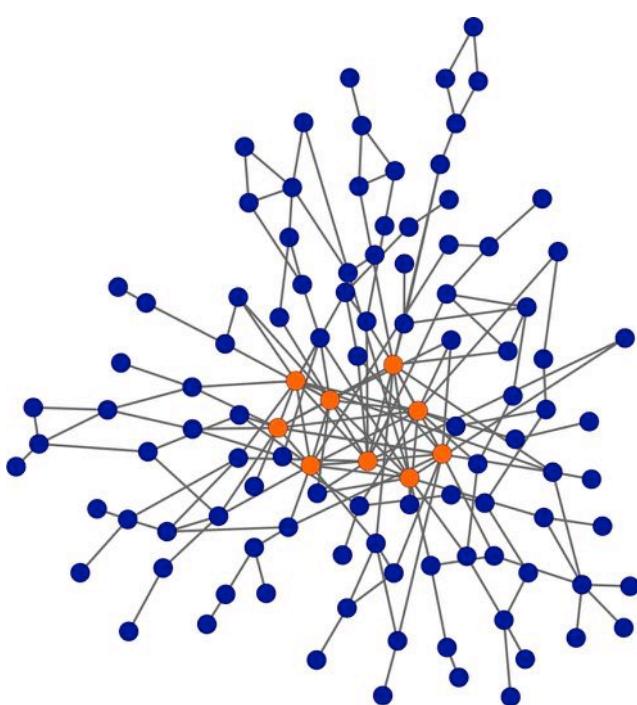



Figura 24-6: Ilustração de uma rede de livre escala consistindo de 109 proteínas, na qual o grau de distribuição segue uma lei de potência. Neste tipo de rede, as proteínas *hubs* (nós laranjas) tem papel essencial na manutenção da integridade da rede.

Diferentemente da rede aleatória que apresenta um número fixo de N nós, as redes de livre escala apresentam uma ordem dinâmica de estruturação que permite o crescimento da rede pela adição de novos nós. Assim, a rede aleatória consiste de um sistema aberto que inicia com um pequeno grupo de nós e aumenta de tamanho exponencialmente no tempo devido à inserção de novos nós. A probabilidade deste novo nó se conectar a nós com grande número de conexões é maior, sendo chamada de conexão preferencial. Por exemplo, imagine que você está buscando um artigo sobre determinado assunto na Internet. Certamente os artigos que você encontrará mais facilmente serão publicações com alto grau de conexão por serem mais conhecidos e bem citados quando comparadas a publicações pouco citadas e, consequentemente, menos conhecidas.

Estes dois mecanismos, crescimento da rede e conexão preferencial originaram o algoritmo do modelo Barabási-Albert, que estabelece que o crescimento inicia-se como uma pequena rede, sendo que a cada instante de tempo um novo nó com m conexões é adicionado, onde a probabilidade do novo nó se conectar ao nó i que está previamente presente depende de k_i (grau de i):

$$\prod(ki) = \frac{k_i}{\sum_j k_j}$$

Esse crescimento gera uma rede de livre escala com expoente de grau $\gamma = 3$. Após t instantes de tempo, temos uma rede com $N = t + m_0$ e m_t conectores.

As características da rede de livre escala a tornam uma rede que apresenta um pequeno número de nós altamente conectados (*hubs*), o que frequentemente determina suas propriedades. Como já mencionado, falhas na rede (ou remoção de nós aleatórios) apresentam poucas consequências, enquanto que o ataque aos nós altamente conectados tornará a rede fragmentada. Em sistemas biológicos, uma rede bioquímica apresenta alta resiliência contra mutações aleatórias, enquanto que os *hubs* podem ser usados como candidatos importantes para alvo de fármacos. Um exemplo disso seria a proteína EF-Tu. Esta proteína tem papel essencial durante a elongação da síntese proteica, sendo inibida pelo antibiótico quirromicina, que impede que o complexo EF-Tu-GDP seja liberado do ribossomo.

Rede Hierárquica

Como já vimos anteriormente, uma rede pode ser avaliada pelo grau de agrupamento (clusterização) de seus nós. Na maioria das redes baseadas em um sistema real (chamadas de redes reais), como por exemplo, parte de uma via metabólica, o coeficiente de clusterização é significativamente maior se comparado a redes aleatórias. Da mesma forma, ocorre a coexistência da propriedade de livre escala e clusterização nas redes reais, como redes metabólicas e de interação proteica. Contudo, grande parte dos modelos propostos para representar estas redes não consegue descrever a livre escala e a clusterização simultaneamente.

Adicionalmente, muitas redes reais

apresentam módulos, ou seja, a rede é composta de subredes funcionalmente separáveis. Esses componentes separáveis apresentam densa conectividade entre os seus próprios nós, com conectividade mais dispersa em relação a componentes de outros módulos. Isso ocorre porque cada módulo apresenta a capacidade de executar uma tarefa identificável, diferente de outro módulo. Contudo, essa "separação" de tarefas não significa que um módulo é independente de outro, mas sim que tem funções distintas.

Dessa forma, é necessário combinar a propriedade de livre escala, o alto grau de agrupamento e a modularidade de uma forma interativa, gerando a rede hierárquica. A estrutura hierárquica é convencionalmente representada por um dendrograma ou uma árvore e atua relacionando os nós mais próximos na rede, conforme Figura 25-6. Essas redes podem ser formadas basicamente pela duplicação de *clusters* e repetidas indefinidamente, integrando uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independentes do tamanho do sistema. Muitas vezes, em redes reais, a modularidade não apresenta um limite claro, sendo reconhecida principalmente por nós altamente conectados entre si e conectados a outros módulos.

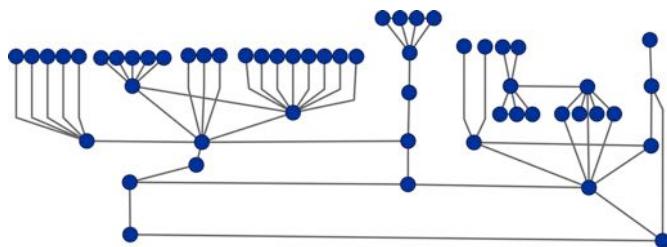



Figura 25-6: Ilustração de uma rede hierárquica consistindo de 55 proteínas em modelo de dendrograma onde é possível observar sua modularidade intrínseca.

A principal característica dessas redes que não é compartilhada por redes aleatórias ou de livre escala é a hierarquia intrínseca, sendo representada também na sua arquitetura. Essa característica hierárquica pode ser, ainda, analisada quantitativamente, como observado por Dorogovtsev e colaboradores em 2002, que construíram um gráfico de livre escala determinístico, na qual o coeficiente de clusterização de um nó que possui k conexões segue a lei de escala $C(k) \sim k^{-1}$. Portanto, o modelo de rede hierárquico integra uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independente do tamanho do sistema.

6.6. Perturbação e conectores

Como visto anteriormente, um grafo consiste de um conjunto de nós e um conjunto de conectores que conectam esses nós. Portanto, os nós são as entidades de interesse e os conectores representam as relações entre as entidades.

Quando tratamos de sistemas biológicos, podemos levar em consideração diferentes entidades como, por exemplo, DNA, RNA, metabólitos, pequenas moléculas e/ou proteínas. Estes componentes biológicos não atuam isoladamente, mas sim dependem da interação com outros componentes. Para que ocorra essa interação (comunicação) é necessária a presença de conectores.

Conectores podem ser interações físicas, bioquímicas ou funcionais. Por exemplo, em redes metabólicas, conectores podem ser reações que convertem um metabólito em outro ou enzimas que catalisam essas reações; em redes de regulação gênica, conectores podem representar a ligação física de um fator de transcrição nos elementos regulatórios; em redes de doenças, conectores podem representar as mutações genéticas associadas à doença; e em redes proteicas, os conectores podem ser ligações físicas entre as proteínas.

Como apresentado anteriormente, as redes podem ser direcionadas e não direcionadas. Esse comportamento da rede depende da natureza da interação e, obviamente, da direcionalidade dos conectores (Figura 26-6). Em redes direcionadas, a interação entre dois nós tem uma direção bem definida que representa, por exemplo, a direção do fluxo do substrato ao produto em uma rede metabóli-

ca. Em redes não direcionadas, a ligação não tem uma direção definida, tal como a interação física entre proteínas.

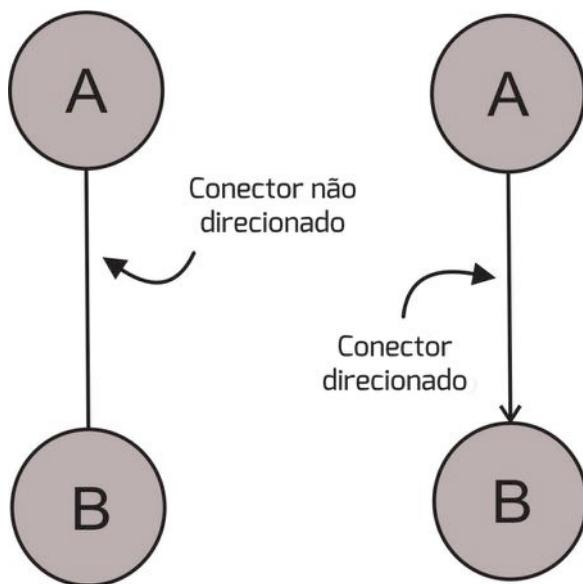



Figura 26-6: Representação de um conector não direcionado e um direcionado.

Na abordagem da biologia de sistemas tão importante quanto conhecer os nós que interagem entre si em uma rede é compreender, por exemplo, que tipo de interação pode ocorrer na rede em questão, quais conectores são mais relevantes à rede e qual o impacto da perturbação de um conector. Nesta seção iremos discutir os tipos de conectores entre diferentes componentes de uma rede envolvendo proteínas e as consequências da ruptura nestas conexões.

Interação proteína-proteína

A interação proteína-proteína é comum e crucial a vários processos celulares, tais como na ligação enzima-inibidor e na interação antígeno-anticorpo. Os diferentes tipos de complexos proteicos têm sido definidos na literatura como obrigatórios e não obrigatórios. No complexo obrigatório, as proteínas não podem funcionar separadamente, diferindo do complexo não obrigatório onde as proteínas associam-se e dissociam-se dependendo de fatores externos, podendo também exercer funções fora do complexo.

De acordo com a estabilidade e o mecanismo de formação do complexo, incluindo o tipo de conexão entre as proteínas, as interações podem ser conceitualmente separadas em dois grupos: aquelas que são permanentes e aquelas que são temporárias. E, embora não exista um limite bem definido para essa separação, tendências têm sido observadas em relação a suas propriedades biológicas (Figura 27-6).

Em relação à estrutura, por exemplo, interações temporárias são caracterizadas por interfaces proteicas pequenas, enquanto que as interfaces de proteínas interagindo permanentemente são maiores. Consequentemente, complexos proteicos com interfaces maiores tendem a apresentar um maior grau de mudança conformacional após a ligação. Além disso, componentes de complexos permanentes tendem a ser co-expressos e mais estáveis. Esta estabilidade gera uma pressão seletiva maior e em função disso, uma taxa evolutiva mais lenta.

Como será discutido adiante, interação transitória tende a ser *date*, isto é, as proteínas podem se conectar em diferentes tempos e a interação permanente tende a ser *party*, isto é, conexão proteica forte e constante.

As proteínas com conectores permanentes existem somente em sua forma complexada e são muito estáveis, enquanto aquelas com conectores transitórios possuem a capacidade de associação e dissociação *in vivo*. Dentre as proteínas com conectores transitórios, há aquelas em que a associação/dissociação é resultante de uma conexão com baixa afinidade, porém constante (interações temporárias fracas) e aquelas em que a associação/dissociação é desencadeada por um processo ativo (interações temporárias fortes) como, por exemplo, uma mudança conformacional ocorrida em consequência de um fator ligante.

A diferença entre as interações acima citadas é distinguida puramente pelas propriedades da estrutura da interface proteica, isto é, da superfície de contato das proteínas. Essas propriedades conferem afinidade e especificidade, e são determinadas principalmente por forças intermoleculares como comple-

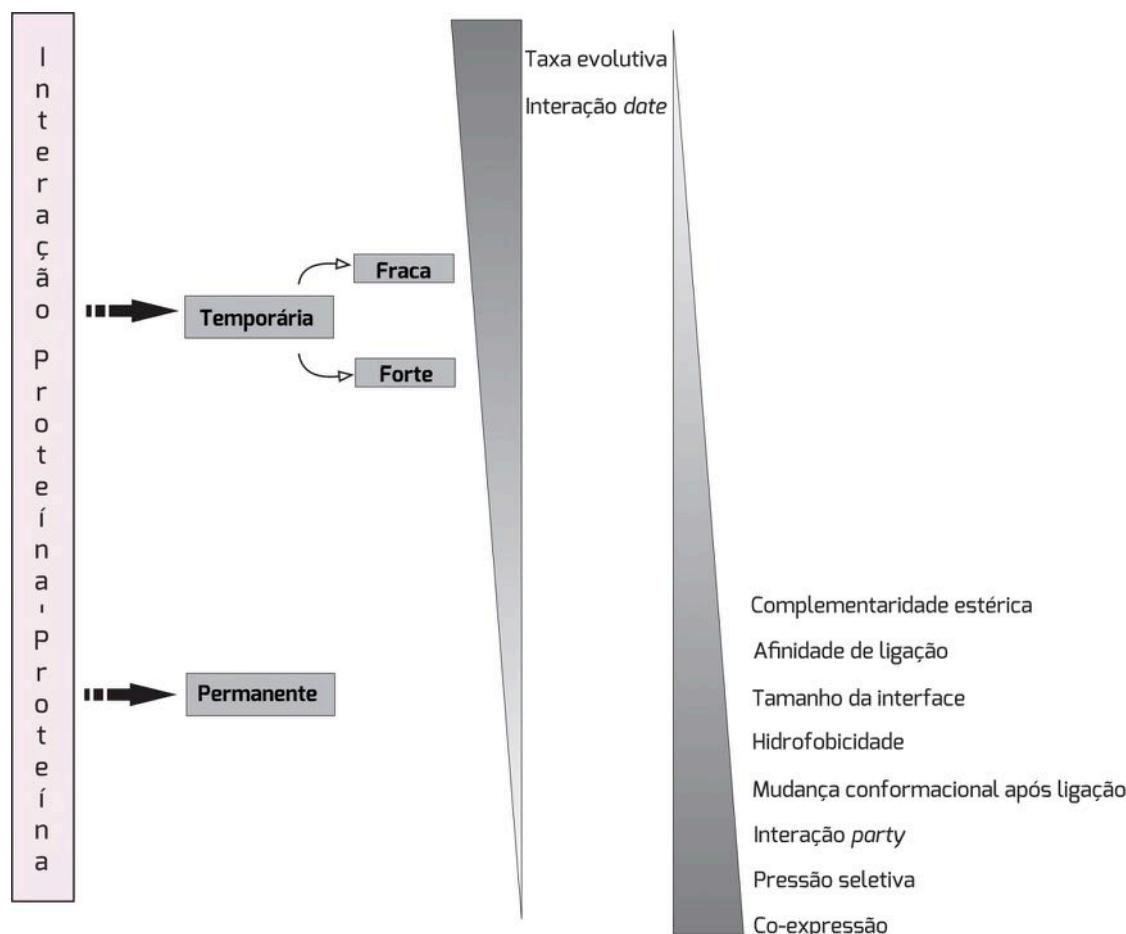



Figura 27-6: Modelo esquemático representando os diferentes tipos de interações proteína-proteína e as propriedades biológicas relacionadas. Quanto maior o tamanho da base e a intensidade da cor do triângulo, maior é a relação entre o modo de interação proteica e a propriedade biológica.

mentaridade estérica, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A complementaridade estérica otimiza as interações de van der Waals entre o complexo. Normalmente, estas interações de fraca energia ocorrem em função da polarização transiente de ligações carbono-hidrogênio ou carbono-carbono e, apesar de fracas, são extremamente importantes para o processo de reconhecimento intermolecular pois crescem em intensidade com a área de interação. Complexos com conexões permanentes exibem alta complementaridade estérica nas proteínas em contato, enquanto complexos com conexões temporárias demonstram baixa complementaridade.

Como as interações de van der Waals, as interações hidrofóbicas são pontualmente fracas e ocorrem em função da interação entre cadeias ou subunidades apolares. Os complexos com conexões permanentes normalmente persistem no estado ligado, sendo a força hidrofóbica mais significativa. Já em conectores transitórios, a alta hidrofobicidade se torna desfavorável, pois esses complexos permanecem ligados por menos tempo.

As forças de atração eletrostáticas são aquelas resultantes da interação entre dipolos e/ou íons de cargas opostas e representam força significativa na interação proteína-proteína, podendo definir o tempo de vida do complexo.

Dentre as forças intermoleculares discutidas acima, o fator dominante da interação permanente entre proteínas consiste nas interações hidrofóbicas, enquanto várias forças

participam de interações temporárias entre proteínas. Além disso, proteínas interagindo de forma temporária possuem interfaces que são menores em tamanho do que as interfaces de proteínas permanentes, os aminoácidos que compõem a interface e a proporção de resíduos hidrofóbicos não diferem drasticamente do resto da superfície proteica e as interfaces são levemente ricas em grupos polares neutros e em água.

O tipo de interação também confere graus diferentes de restrição (pressão seletiva) na evolução da proteína. Proteínas com interação permanente tendem a evoluir em uma velocidade menor comparada a proteínas que formam complexos temporários, bem como possuir pressão seletiva maior e menor plasticidade em sua sequência.

Evidências sugerem que o modelo duplicação-divergência aplica-se à evolução das redes proteicas. Uma das predições é que na duplicação das proteínas algumas ou todas as conexões podem ser herdadas da proteína ancestral. Consistente com esta hipótese, proteínas parálogas tendem a compartilhar padrões de interação em uma frequência maior do que a esperada ao acaso. No entanto, tem sido proposto que depois que a duplicação gênica ocorre, as interações entre as proteínas são rapidamente perdidas. Portanto, duplicações recentes são mais prováveis de compartilhar interações, comparadas a duplicações mais ancestrais.

Outra distinção acerca da interação proteica refere-se à interação funcional e interação física. A interação funcional pode ou não corresponder a uma interação física direta em algum processo biológico. Assim, na interação física, a proteína A conecta-se a proteína B e, na interação funcional, a proteína A atua com a proteína B. Como exemplo de interação funcional podemos imaginar dois produtos gênicos que interagem em uma mesma via em um processo biológico, mas não se conectam fisicamente.

O tipo de interação tem um papel importante na determinação do comportamento das proteínas. Como já vimos, *hubs* são proteínas envolvidas em um grande número de interações (altamente conectadas) dentro de uma rede proteica. Algumas proteínas *hub* são altamente co-expressas com outras proteínas do módulo, o que implica na existência de complexos estáveis (permanentes). Outras proteínas possuem expressão independente, sugerindo a ligação com proteínas em diferentes tempos, de modo transitório. Esses *hubs* são classificados como *party* e *date hubs,* respectivamente.

Na construção de redes proteicas, a diferenciação entre complexos permanentes e transitórios tem importantes implicações. Por exemplo, na prospecção de novos fármacos, a alteração do padrão de interação entre proteínas temporárias por modulação farmacológica ocorre mais facilmente em comparação a proteínas que formam complexos permanentes. Portanto, uma rede de interação proteica não é um processo estático, mas sim corresponde a um constante fluxo de informações. Por conseguinte, na análise de dados de interação proteína-proteína a discriminação das características da interação e/ou o uso de centralidades de conectores é fundamental para obter modelos mais realísticos.

Interação proteína-ácidos nucleicos

Proteínas que se ligam a ácidos nucleicos têm um papel central em todos os processos regulatórios que controlam o fluxo de informação genética. Por exemplo, proteínas podem inibir, ativar e coordenar a transcrição do DNA, auxiliar e manter o empacotamento e o rearranjo do DNA e o processamento do RNA, coordenar a replicação do DNA, promover a síntese de proteínas e sinalizar o reparo do DNA, entre outros.

Esses possíveis papéis fisiológicos são determinados pela afinidade e especificidade da interação DNA-proteína, que é a habilidade da proteína em distinguir seu sítio de ligação do restante do DNA. Estas propriedades dependem de interações precisas entre a sequência de aminoácidos da proteína e os nucleotídeos do sítio específico de ligação do DNA.

As proteínas que se ligam a ácidos nucleicos podem ser, de forma simplificada separadas em três grupos de acordo com a função:

i) enzimas, onde a principal função da proteína é modificar a organização do ácido nucleico, como no caso das endonucleases, glicosiltransferases, glicosilases, helicases, ligases, metiltransferases, nucleases, polimerases, recombinases, topoisomerases, translocases e transposases, entre outras;

ii) fatores de transcrição, onde a principal função da proteína é regular a transcrição e a expressão gênica como por exemplo, TFIIA, TFIIB, TFB, entre outros;

iii) proteínas estruturais que ligam-se ao DNA, que têm como principal função suportar a estrutura e a flexibilidade do DNA ou agregar outras proteínas, por exemplo, proteínas centroméricas, proteínas envolvidas no empacotamento e na manutenção/proteção do DNA, proteínas de reparo, proteína envolvidas na replicação e proteínas teloméricas, entre outras.

A interação proteína-proteína também é necessária para uma eficiente interação entre proteínas e ácidos nucleicos. A interação proteína-proteína com o DNA pode ocorrer de três modos de acordo com a direção e o eixo da dupla hélice do DNA (Figura 28-6):

i) a direção da interação entre as proteínas e o eixo da dupla hélice é perpendicular;

ii) a direção da interação da proteína é paralela ao eixo da dupla hélice;

iii) ambos os modos de interação são observados ao mesmo tempo.

Assim como na formação de complexos proteicos, discutido anteriormente, a formação de complexos DNA-proteína ou RNA-proteína também envolve forças intermoleculares, tais como van der Walls, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A região da proteína que reconhece a sequência do ácido nucleico é denominada motivo. Os motivos hélice-volta-hélice, dedo de zinco e zíper de leucina são os mais comuns encontrados nas proteínas que interagem com ácidos nucleicos.

O motivo hélice-volta-hélice é um dos elementos normalmente encontrados nos fatores de transcrição e nas enzimas de procariotos e eucariotos, sendo formado por duas hélices α conectadas por uma volta. O motivo liga-se a cavidade maior do DNA e, em muitos complexos, o contato direto é feito entre a cadeia de aminoácido e a sequência de bases do ácido nucleico.

Já o motivo dedo de zinco é encontrado principalmente em fatores de transcrição de eucariotos. Um dedo de zinco é composto por duas folhas β antiparalelas e uma hélice α, sendo o íon zinco fundamental para garantir a estabilidade deste tipo de domínio. Subunidades proteicas contêm múltiplos dedos de zin-

Figura 28-6: Modos de interação proteína-proteína com a dupla hélice do DNA. A) perpendicular; B) paralela e C) ambas as direções são observadas.

co que se enrolam no DNA formando uma espiral, inserindo a hélice α na cavidade maior do DNA.

Fatores de transcrição de eucariotos e procariotos também podem conter o motivo zíper de leucina, encontrado em proteínas regulatórias. Esse motivo é formado por duas hélices α paralelas, unidas por resíduos de leucina.

A estrutura do zíper de leucina pode ser dividida em duas partes: a região de dimerização e a região de ligação ao DNA. A dimerização é mediada pela formação de uma estrutura enrolada na região carboxi-terminal de cada hélice com sete resíduos de leucina. A região que se liga ao DNA, também conhecida como região básica, é encontrada na região amino-terminal da hélice que se projeta na cavidade maior do DNA. Embora motivos de diferentes famílias de DNA sejam similares estruturalmente, pouca homologia é observada fora do motivo. Há baixa identidade entre motivos de diferentes famílias de proteínas e esta variação permite, portanto, o reconhecimento de diferentes conjuntos de sequências de DNA. Além disso, a posição do domínio dentro da cavidade maior do DNA também varia, refletindo a necessidade funcional e estrutural de cada proteína.

A afinidade e a especificidade na ligação de proteínas ao DNA não podem ser endereçados somente a alguns resíduos de aminoácidos, mas o envolvimento de toda a proteína deve ser considerado. Por exemplo, a maioria das proteínas que se ligam ao DNA possuem domínios desordenados que contribuem para o reconhecimento do DNA em vários níveis.

Proteínas com domínios desordenados são proteínas que não apresentam estrutura $2^{ária}$ e $3^{ária}$ sob condições fisiológicas e na ausência de ligantes naturais. Essas proteínas possuem alta especificidade e baixa afinidade na interação, são capazes de interagir com mais de uma proteína e alvos de modificações pós-traducionais, possuindo a capacidade de manter sua função mesmo em ambientes extremos. Na interação com o DNA, o domínio desordenado da proteína não é crucial à formação do complexo, mas pode influenciar o reconhecimento da sequência do DNA, conferindo seletividade e afinidade de ligação.

Além da característica das cavidades na molécula de DNA, da presença de motivos específicos nas proteínas ou ainda da ocorrência de domínios desordenados, outros fatores podem influenciar a interação do DNA-proteína, tais como a flexibilidade e a afinidade da proteína pelo DNA e presença de água no meio.

Muitas proteínas são flexíveis ao ponto de alterar sua conformação quando se ligam ao DNA, enquanto outras são conhecidas por alterar a conformação do DNA após a ligação. A afinidade da interação entre o DNA e uma proteína tende a estar relacionada à relevância funcional da proteína. Por exemplo, a afinidade de um fator de transcrição por seu sítio de ligação é proporcional à ativação que ele exerce. Ainda, alguns contatos mediados por água foram observados entre proteínas e o DNA, participando de redes de ligações de hidrogênio que conferem estabilidade ao complexo.

Interação entre proteínas e pequenos compostos

Considerando-se que a interação proteína-proteína normalmente envolve superfícies relativamente grandes, pode-se imaginar que moléculas menores não seriam efetivas na modulação da ligação dos complexos por apresentarem áreas menores e, por conseguinte, interações menos intensas. Contudo, ao empregarmos estruturas químicas diferentes de aminoácidos, podemos não só compensar esta redução na área de contato mas produzir moléculas com afinidade maior do que os próprios ligantes fisiológicos envolvidos do processo de interesse.

Adicionalmente, estas moléculas de baixa massa molecular tendem a apresentar muitas vantagens terapêuticas em relação a proteínas, dentre as quais se destaca sua maior estabilidade metabólica e consequente maior biodisponibilidade. Podem atuar diretamente – via inibição da interface proteína-proteína – ou indiretamente – via ligação a um sítio alostérico que induz uma mudança conformacional do alvo da proteína ou da molécula associada.

A busca de novos fármacos deve levar em conta o tipo de complexo proteico alvo. A formação de complexos permanentes pode ser considerada uma continuação do enovelamento da proteína, sendo o dobramento fi-

nal das subunidades parte deste processo. Assim, esse tipo de complexo é menos propenso à modulação farmacológica, sendo mais interessante explorar o processo de dobramento em si como alvo de pequenos compostos. Já as interfaces das proteínas de complexos temporários são alvos efetivos ao planejamento de novos moduladores terapêuticos.

Para que pequenas moléculas modulem a interação proteica, estratégias têm sido estabelecidas e dois principais mecanismos do controle regulatório têm sido utilizados: a inibição e a estabilização (Figura 29-6). Das estratégias mais exploradas, destaca-se a inibição da interação proteína-proteína.

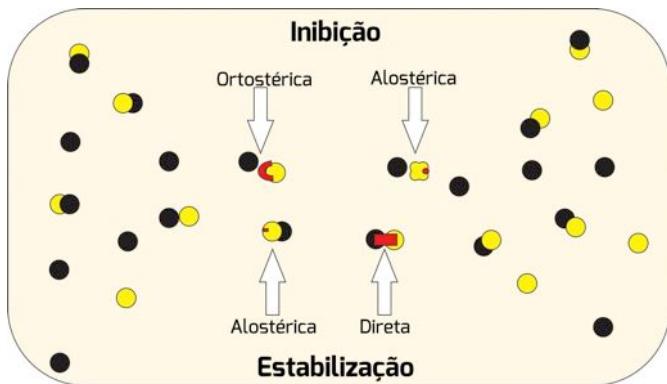



Figura 29-6: Dois principais mecanismos de modulação da interação proteína-proteína utilizando pequenos compostos. Diferentes proteínas são apresentadas em preto e amarelo. Pequenos compostos são apresentados em vermelho.

O modo de ação da maioria dos inibidores de interação proteica é baseado na ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos *hot spots* críticos da interface e competindo com a proteína original. Esse tipo de inibição é conhecido como ortostérica. Na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante (Figura 29-6).

Pequenas moléculas estabilizadoras da interação proteína-proteína também demonstram dois modos gerais de ação. Primeiro, um estabilizador pode ligar-se a uma única proteína, na qual aumenta a afinidade de ligação mútua das proteínas do complexo de um modo alostérico. Segundo, a molécula estabilizadora liga-se à superfície do complexo proteico, fazendo contato com ambas as proteínas ligantes e aumentando a afinidade de ligação mútua entre elas. Assim, a inibição estabilizadora pode ser denominada alostérica (ligada a uma proteína) ou direta (ligada ao menos a duas proteínas).

A ativação por pequenos compostos é, normalmente, um processo mais intrincado pois, além da ligação, é necessário o correto desencadeamento da cascata de ativação. Compostos que induzem a interação proteica são chamados de dimerizadores. Inúmeras vias de sinalização celular iniciam a partir da dimerização proteína-proteína. A principal ideia do uso de dimerizadores é a indução de interação entre duas proteínas por pequenas moléculas que levam à ativação da via de sinalização celular. Na literatura científica foi observado que dimerizadores podem induzir proliferação celular, transcrição e apoptose.

Perturbação dos conectores

Perturbações podem ocorrer em todos os sistemas, e em sistemas biológicos não é diferente. Nos interatomas, essas perturbações podem variar desde a remoção de um ou mais nós até a remoção de conectores. Desta forma, as consequências na estrutura e na função do sistema irão diferir drasticamente dependendo do tipo de perturbação ao qual a rede foi exposta. Como exemplo, podemos imaginar uma rede de proteínas que confere um fenótipo específico (Figura 30-6).

A remoção do nó não somente incapacita a função deste, mas também a de outros nós, causando a ruptura nas vias de todos os nós vizinhos. Uma perturbação no conector, que remove uma ou poucas interações mas deixa o restante da rede intacta e funcionando, pode ter efeitos mais sutis no sistema, não necessariamente alterando o fenótipo. Contudo, a consequência do desarranjo da rede após a remoção de nós ou de conectores depende da importância do nó e do conector à rede. Essas informações de conectores e nós

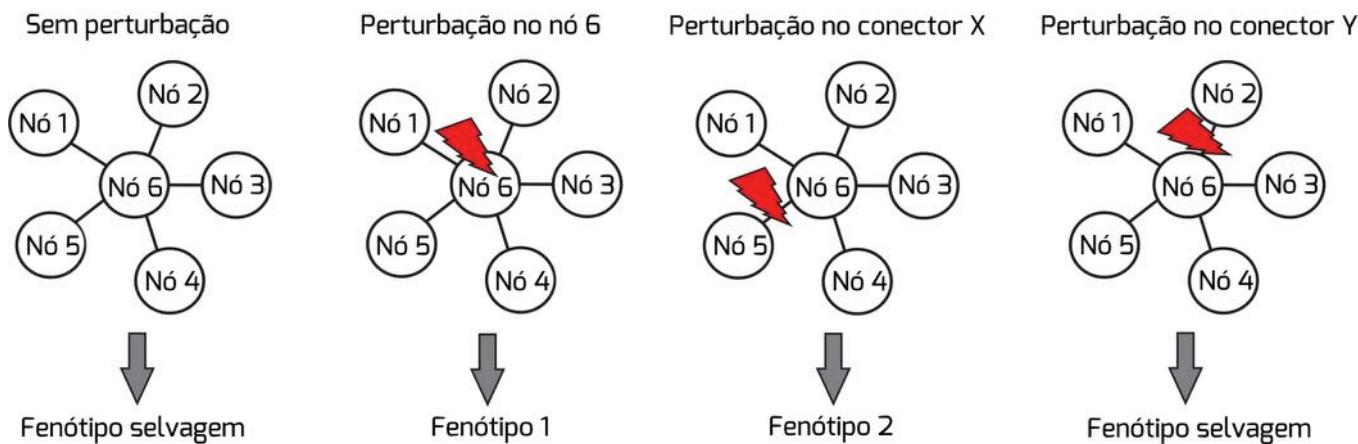



Figura 30-6: Rede hipotética de proteínas relacionada a um fenótipo específico representando diferentes tipos de perturbação e suas consequências. Neste exemplo o nó 5 e o conector entre os nós 5 e 1 são essenciais à manutenção do fenótipo selvagem.

mais informativos de uma rede podem ser obtidas, por exemplo, pela análise da resiliência e percolação da rede, vista anteriormente.

A distinção entre modelos de remoção de nó e perturbação de conectores - alteração interação-específica e conector-específica (*edge-specific* ou "*edgetic*"), respectivamente - pode providenciar novas pistas nos mecanismos básicos de doenças humanas, tais como diferentes classes de mutações que levariam a modos dominantes ou recessivos de herança genética.

Em uma rede proteica, a remoção de um nó pode representar a remoção de uma proteína, causado por uma mutação crítica no gene que desestabiliza a estrutura da proteína. Já a remoção de um conector pode representar uma mudança específica em distintas interações bioquímicas e biofísicas, preservando certos domínios da proteína.

Em relação a genes envolvidos em múltiplas doenças, foi demonstrado que alelos *edgetic* responsáveis por diferentes doenças consistem em distintas perturbações *edgetic* que, por sua vez, tendem a estar localizados em diferentes domínios de interação proteica, conferindo fenótipos diferenciados.

Pesquisadores analisaram cerca de 50.000 alelos mendelianos associados a doenças genéticas hereditárias e observaram que aproximadamente a metade foi potencialmente *edgetic*. Nesta análise foram consideradas deleções e mutações truncadas dentro dos domínios da proteína que grosseiramente desestabilizaram a estrutura da proteína, como remoção de nó, mutações com alteração em quadro de leitura que afetaram sítios de ligação específicos e mutações truncadas que preservaram certos domínios da proteína como perturbação *edgetic*. Alelos truncados foram menos propensos a expressar proteínas estáveis em comparação a alelos que alteraram o quadro de leitura, podendo diferir doenças hereditárias mendelianas envolvendo remoção de nó *versus* perturbação *edgetic*.

Um alelo *edgetic* pode ser identificado pela falta de um subconjunto de interações, quando possuem defeitos nas interações provavelmente devido a mudanças específicas dentro ou próximo a sítios de ligação da proteína ou quando fenótipos *in vivo* diferem daqueles causados por perturbações nulas (genótipos nulos).

Dependendo da rede, o fenômeno de perturbação de um único conector pode ser mais provável do que da remoção de um nó. Dependendo do conector rompido, o impacto à rede pode ser maior, pois diferentes conectores (interações) têm diferentes níveis de importância (vulnerabilidade). Conectores com alto valor de *edgebetweenness* podem causar fragmentação da rede em componentes desconectados, caso sejam rompidos, como por exemplo no caso de conectores entre *clusters*. Esse tipo de conector é assim chamado de *cut-edge*. Já conectores com baixo valor de *edgebetweenness*, quando eliminados da rede, podem ser substituídos por vias alternativas, como por exemplo no caso de

conectores dentro de *clusters*. Assim, conectores *interclusters* tendem a ser mais vulneráveis quando comparados aos conectores *intraclusters* em uma determinada rede.

6.7. Conceitos-chave

Assortatividade: tendência de nós interagirem com nós similares a eles mesmos.

Betweenness: parâmetro que estima a relação entre dois nós, ou seja, leva em consideração a quantidade de caminhos mais curtos que passam entre eles.

Biologia de sistemas: área da bioinformática que estuda sistemas moleculares complexos e como as moléculas interagem entre si.

Caminho: sequência consecutiva de nós em um grafo sem repetições, estando cada nó adjacente interligado por um conector.

Caminho geodésico: definido pela via mais curta dentro de uma rede entre dois nós quaisquer.

Circuito: sequência de nós sem repetição com um conector entre cada par de nós adjacentes na sequência, onde o nó inicial coincide com o nó final.

Clique: é definido como um grafo com alta conectividade entre seus elementos integrantes. Sendo assim, clique também é considerado um sinônimo de *cluster*.

Closeness: valor que indica os caminhos mais curtos entre um nó n e todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó.

Complexo proteico: grupo de proteínas formado pela associação de duas ou mais cadeias polipeptídicas.

Comprimento do caminho: definido pelo número de conectores que definem o caminho, ou então, pelo número de nós da sequência menos um.

Conector *Cut-edge*: conector que quando rompido causa fragmentação da rede.

Date hubs: são hubs que se ligam a diferentes proteínas em diferentes módulos (intermódulo), ou seja, diferente tempo e/ou espaço, consequentemente, apresentado um papel global na rede.

Desassortatividade: tendência de nós interagirem com nós diferentes deles mesmos.

Diâmetro: indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo.

Dimerização: corresponde à união de dois monômeros, formando um dímero. Ou seja, é a formação de uma molécula a partir de duas moléculas menores.

Dimerizadores: compostos que induzem a dimerização, neste caso a interação proteica.

Distribuição de Poisson: distribuição aplicada a probabilidade de ocorrência de um evento em determinado intervalo de tempo.

Edgebetweenness: parâmetro que indica o número de caminhos mais curtos entre pares de nós que percorrem um determinado conector.

Edgetic: perturbação causada em um conector específico, portanto em uma interação específica na rede.

Forças intermoleculares: forças que mantêm as moléculas unidas durante a interação.

Gargalo (*bottleneck*): proteína que apresenta alto grau de *betweenness*.

Grau de nó (*node degree*): parâmetro referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó.

Hipergrafo: rede caracterizada pela presença de hipervértices.

Hipervértices: Conectores que interligam nós que apresentam propriedades distintas nos hipergrafos.

Hot spot proteico: locais essenciais da interface com alta afinidade de ligação.

Inibição alostérica de uma proteína: na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante.

Inibição ortostérica de uma proteína: inibição causada pela ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos *hot spots* críticos da interface e competindo com a proteína original.

Interface proteica: área através da qual as macromoléculas se comunicam e exercem sua funcionalidade.

Modularidade (clusterização): padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados.

Multiconector, interações: quando há dois ou mais conectores ligando os mesmos nós na rede em redes direcionadas.

Multidígrafo: rede direcionada com a presença de multiconectores.

"Mundo pequeno", efeito: define que existe um caminho mínimo entre um nó de origem e um nó de destino.

Ontologia gênica: tipo de análise que tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um mesmo processo biológico.

Party hubs: proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligação no mesmo tempo e/ou espaço.

Pleiotrópico, efeito: proteínas pleiotrópicas são aquelas que apresentam múltiplos efeitos em um sistemas biológico.

Rede: representação gráfica da interação entre nós por meio de vértices.

Rede bipartida: existe uma partição da rede, por exemplo, partição A e partição B, sendo os nós presentes na partição A adjacentes apenas a nós da partição B, e vice-versa.

Rede direcionada: apresentam conectores que orientam o fluxo da informação em uma direção.

Rede não direcionada: os conectores desta rede não apresentam uma direção orientada.

Rede ponderada: são redes que se caracterizam pela presença de atributos associados a conectores e nós.

Resiliência: capacidade de uma rede a tolerar a deleção de seus nós por falha ou ataque.

Taxa evolutiva: medida das mudanças ocorridas numa entidade (gene, proteína, organismo, população) evolutiva ao longo do tempo.

Teoria da Percolação: tem por objetivo investigar o comportamento das propriedades de conectividade de uma rede.

Topologia de redes: estrutura e disposição de conexões entre os nós.

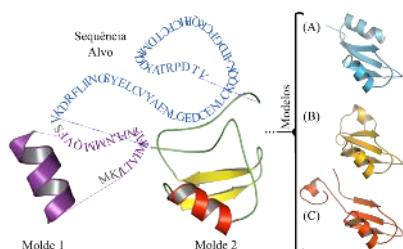
Vulnerabilidade do conector: grau de importância do conector.

6.8. Leitura recomendada

BARABÁSI, Albert-László; OLTVAI, Zoltán N. Network biology: understanding the cell's functional organization. ***Nat. Rev. Genetics***. 5, 101-113, 2004.

GURSOY, Attila; KESKIN, Ozlem; NUSSINOV, Ruth. Topological Properties of Protein Interaction Networks from a Structural Perspective. ***Biochem. Soc. Trans***. 36, 1398-1403, 2008.

LEVY, Emmanuel D.; PEREIRA-LEAL, Jose B. Evolution and Dynamics of Protein Interactions and Networks. ***Cur. Op. Struct. Biol.*** 18, 1–9, 2008.

MASON, Oliver; VERWOERD, Mark. Graph theory and networks in Biology. ***IET Systems Biol***. 1, 89-119, 2007.

NEWMAN, Mark E. J. The structure and function of complex networks. ***SIAM Rev***. 45, 167-256, 2003.

YU, Haiyuan; et al. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. ***PLoS Comp. Biol***. 3, e59, 2007.

WAGNER, Günter P.; PAVLICEV, Mihaela; CHEVERUD, James M. The road to modularity. ***Nat. Rev. Genetics***. 12, 921-931, 2007.

7. Modelos Tridimensionais

Geração de múltiplos modelos para a estrutura de uma determinada sequência de aminoácidos.

Priscila V. S. Z. Capriles
Raphael Trevizani
Gregório K. Rocha
Laurent E. Dardenne
Fabio Lima Custódio

7.1. Introdução

O rápido avanço na computação científica verificado na última década, principalmente quanto ao aumento da capacidade de processamento dos computadores a custos relativamente baixos, tem permitido que classes importantes de problemas científicos na área da bioinformática, no estudo de biomoléculas e sistemas biológicos, possam ser abordadas com cada vez mais sucesso. Dentre estas áreas, a predição de estruturas tridimensionais de proteínas destaca-se pela sua importância, o que tem atraído um grande número de pesquisadores ao redor do mundo. Um exemplo deste interesse está na criação de um encontro bianual de caráter mundial, intitulado CASP - *Critical Assessment of Protein Structure Prediction*, com o objetivo de avaliar o estado da arte da capacidade de predição de diferentes metodologias desenvolvidas.

A predição de estruturas tridimensionais de proteínas se caracteriza por possuir aplicações práticas de grande impacto terapêutico e biotecnológico. Está diretamente relacionada a múltiplas áreas da bioinformática e modelagem molecular, tais como o atracamento proteína-ligante (ver capítulo 9), aplicado ao desenho racional de fármacos baseado em estruturas, o desenho de novas proteínas com funções específicas (nanotecnologia e engenharia de proteínas) e a própria elucidação de estruturas a partir de dados experimentais, por exemplo, de ressonância magnética nuclear (RMN). Avanços teóricos e metodológicos implicariam em impactos diretos na saúde e no bem estar da sociedade. No entanto, apesar dos avanços realizados nos últimos anos, o desenvolvimento de metodologias capazes de alcançar um elevado grau de previsibilidade e acurácia continua sendo um importante desafio.

7.2. Estrutura 3D de proteínas

Proteínas

A função de uma proteína está intima-

mente associada à sua estrutura tridimensional. Essa é a afirmativa fundamental que inspira todas as buscas por um método que seja capaz de prever a estrutura nativa de uma proteína a partir da sua sequência de aminoácidos. Tal método poderia ajudar na compreensão e no melhor aproveitamento do potencial contido na grande quantidade de informação biológica, na forma de sequências, que vem sendo gerada graças ao sucesso dos projetos genoma.

"As informações sobre a estrutura de uma proteína estão armazenadas em uma sequência codificada nos genes de um organismo". Assim diz um dos principais paradigmas da biologia, postulado por Anfinsen em 1973. A sequência é traduzida através de um complexo aparato celular em uma estrutura tridimensional funcional. Entender todos os mecanismos e forças por traz desse processo seria um enorme avanço científico que influenciaria praticamente todas as áreas das ciências da vida. Esse produto funcional da tradução, chamado de estrutura nativa, é uma macromolécula estável, em condições fisiológicas, formada por ligações peptídicas entre os aminoácidos.

Apesar de estável, a estrutura nativa está longe de ser uma molécula estática. Trata-se de uma estrutura flexível, com movimentos específicos, muitos dos quais são diretamente responsáveis pela função da proteína. Por esse motivo, consideramos o "estado nativo" de uma proteína não como uma estrutura estática, mas como um conjunto de conformações (também chamadas de configurações) de baixa energia livre e biologicamente relevantes que a cadeia assume regularmente no meio no qual exerce suas funções.

Determinação experimental

As principais técnicas para a determinação experimental da estrutura tridimensional de macromoléculas biológicas serão apresentadas nos capítulos 12 e 13. Brevemente, o processo para a obtenção da estrutura tridimensional de uma proteína via técnica de cristalografia por difração de raios-X é composto basicamente pela produção e purificação da proteína alvo, cristalização, coleta e processamento dos dados, resolução da estrutura (empregando informações sobre a sequência de aminoácidos e diferentes programas) e refinamento da estrutura.

A técnica de RMN também requer o conhecimento da sequência de aminoácidos. Contudo, não é necessário que a proteína esteja em um estado de cristal ordenado. A vantagem da RMN é que a estrutura a ser determinada pode estar em solução, apesar de requerer que a proteína solubilizada esteja em altas concentrações. Infelizmente, esta técnica ainda está limitada a proteínas de tamanhos pequenos a médios, limitação não observada para a cristalografia. Mesmo assim, a RMN destaca-se ao revelar informações sobre o comportamento dinâmico das estruturas, incluindo mudanças conformacionais e interações com outras moléculas.

Na RMN, um forte campo magnético alinha os momentos magnéticos dos núcleos atômicos de isótopos que possuem *spin* nuclear diferente de zero (tais como ^{1}H, ^{13}C, ^{15}N, ^{9}F e ^{31}P). Uma fonte de radiofrequência de energia variável é emitida, podendo ser absorvida pelos núcleos atômicos invertendo o alinhamento do *spin* nuclear em relação ao campo magnético externo aplicado. Neste momento, parte da energia é absorvida e o espectro de absorção resultante fornece a informação sobre a identidade do núcleo e seu ambiente químico na vizinhança. Dados de sucessivos experimentos são coletados e um espectro de RMN é gerado contendo as informações sobre todos os deslocamentos químicos de todos os isótopos analisados na proteína.

7.3. Enovelamento de proteínas

O enovelamento de proteínas é objeto de grande interesse de diversas áreas do conhecimento, como mencionado acima. Dada a presença marcante das proteínas em inúmeros processos biológicos, é surpreendente que ainda hoje se saiba tão pouco de como o enovelamento ocorre, permitindo que as proteínas adotem sua estrutura nativa. Estudos sobre o enovelamento de proteínas tratam do processo pelo qual a cadeia peptídica sinteti-

zada adota a sua estrutura tridimensional nativa. Eles diferem dos estudos de predição de estrutura de proteínas (PSP – *Protein Structure Prediction*) por estarem mais interessados no "como" e não no produto final do processo de enovelamento. Mas é justamente este "como" que nos permite conhecer mais detalhes sobre o enovelamento e, a partir destas informações, desenvolver novos métodos de predição de estruturas. De fato, a maioria dos métodos de predição é inspirada em um ou mais aspectos das teorias de enovelamento.

O postulado de Anfinsen e a hipótese termodinâmica

O trabalho laureado de Christian Anfinsen sobre a enzima ribonuclease demonstrou a relação entre a sequência de aminoácidos de uma proteína e sua conformação. A ribonuclease é uma proteína constituída de 124 aminoácidos cuja atividade catalítica é a clivagem de moléculas de RNA. Ela possui em sua estrutura nativa quatro pontes dissulfeto. Sendo estas ligações oriundas da oxidação de resíduos de cisteína espacialmente próximos, podem ser clivadas reversivelmente por um agente redutor.

Anfinsen e seus colaboradores, usaram o reagente denominado β-mercaptoetanol (que forma dissulfetos mistos cistina-β-mercaptoetanol). Em grandes quantidades, este reagente provoca a redução completa de todos os resíduos de cisteína. Contudo, eles notaram que a proteína não podia ser prontamente reduzida a menos que estivesse parcialmente desenovelada por agentes tais como ureia e cloridrato de guanidina. Embora o mecanismo não seja completamente compreendido, esses agentes perturbam as interações não covalentes que estabilizam a estrutura da proteína, provocando o seu desenovelamento.

Quando uma solução da proteína ribonuclease foi incubada com ureia a 8 M e β-mercaptoetanol, observou-se que ela perdia totalmente a sua atividade catalítica. Em outras palavras, a ribonuclease era desnaturada. Isso confirmou a observação de que para que uma proteína exerça a sua função, ela deve estar em sua conformação nativa.

Anfinsen fez então a observação crítica de que a ribonuclease desnaturada, uma vez livre da ureia e do β-mercaptoetanol, por diálise, recuperava lentamente a atividade enzimática. Ele imediatamente percebeu o significado deste achado: os resíduos de cisteína da cadeia eram oxidados pelo ar e a enzima espontaneamente se enovelava para a forma cataliticamente ativa. As experiências de Anfinsen e seus colaboradores mostraram que a informação necessária para especificar a complexa estrutura tridimensional da ribonuclease estava contida em sua sequência de aminoácidos. Estudos posteriores estabeleceram a generalidade desse importante princípio da biologia molecular: a sequência é um importante determinante da conformação proteica.

Em resumo, o postulado de Anfinsen, também conhecido como a hipótese termodinâmica, afirma que, pelo menos para pequenas proteínas globulares, a estrutura nativa é determinada unicamente pela sequência de aminoácidos. Isso equivale a dizer que, nas condições do ambiente (isto é, temperatura, pressão e constituição do solvente) em que o enovelamento ocorre, a estrutura nativa possui três propriedades:

i) A estrutura deve ser única, isto é, uma dada sequência não deve possuir outras conformações com energia livre comparável com a do estado nativo;

ii) A estrutura deve ser estável, isto é, pequenas mudanças no ambiente ao seu redor não devem causar mudanças no enovelamento. Isso leva à imagem de que, pelo menos perto do mínimo global, o enovelamento de proteínas segue um formato de funil, que implicaria na estabilidade do estado nativo;

iii) A estrutura deve ser cineticamente acessível, isto é, o processo pelo qual a forma nativa de uma dada proteína seja atingida deve ocorrer em um tempo compatível com fenômenos biológicos. Proteínas de um único domínio se eno-

velam em uma escala de tempo da ordem de microssegundo até segundos. Para satisfazer esses critérios, durante o enovelamento, a estrutura não deve sofrer mudanças muito bruscas na sua conformação, isto é, movimentos que implicam em barreiras energéticas muito grandes.

Sequências muito diferentes podem adotar estruturas muito parecidas. Ainda, o enovelamento é frequentemente influenciado ou mesmo totalmente dependente de modificações co- ou pós-traducionais, além do ambiente molecular de destino e da participação de chaperonas. Ainda, observou-se que o enovelamento de proteínas em células nem sempre termina na forma nativa, o que levou ao surgimento, durante a evolução, de mecanismos de controle de qualidade do enovelamento proteico.

Origem da estabilidade estrutural

Podemos dizer que as proteínas são estabilizadas pela combinação de interações não covalentes oriundas da interação entre diferentes regiões da cadeia. Nesse contexto, estabilidade se refere à tendência em manter uma conformação nativa. Uma cadeia polipeptídica, em teoria, pode assumir um número muito grande de configurações e, por isso, o estado desenovelado (também chamado de desnaturado) é caracterizado por uma alta entropia conformacional. Essa entropia, juntamente com as interações (por ligações de hidrogênio) com o solvente, leva à estabilização do estado desenovelado.

As interações que contribuem para neutralizar esses efeitos e estabilizar o estado nativo são, além das pontes dissulfeto, interações como ligações de hidrogênio intramoleculares e interações de van der Waals. Note que, para se quebrar uma ligação covalente, é necessário muito mais energia do que para se romper interações não covalentes (aproximadamente 100 vezes mais). E, embora mais fracas, essas interações são muito mais numerosas do que o principal tipo de ligação covalente (pontes dissulfeto) que, em algumas proteínas, estabiliza a estrutura 3^{ária}. Assim, em geral, a conformação com o maior número dessas interações fracas é a configuração de menor energia livre.

Por conseguinte, a estabilidade de uma proteína não é proveniente da simples soma das energias de suas interações não covalentes. Em solução, cada grupo formador de ligações de hidrogênio na cadeia peptídica estava interagindo com moléculas de água antes da estrutura se enovelar. Então, para cada nova ligação de hidrogênio intramolecular formada quando a estrutura se enovela, uma ligação equivalente com o solvente é desfeita. Na prática, um dos principais fatores que impulsionam o enovelamento de uma proteína é o chamado efeito hidrofóbico. Resumidamente, o efeito hidrofóbico pode ser entendido como a tendência de resíduos de aminoácidos hidrofóbicos se agruparem no interior da proteína (que se torna portanto apolar) e dos resíduos hidrofílicos se exporem na superfície da mesma (que se torna portanto polar).

Em soluções aquosas existe uma rede de ligações de hidrogênio entre as moléculas de água. Moléculas do soluto tendem a romper ou atrapalhar a formação dessa rede. Esse efeito é mais pronunciado ao redor de moléculas hidrofóbicas, onde é formada a camada de solvatação (região onde as moléculas de água estão altamente organizadas em um padrão ótimo de formação de ligações de hidrogênio). O aumento da ordenação das moléculas de água na camada de solvatação, ao redor de solutos hidrofóbicos (não-polares) resulta em uma diminuição desfavorável da entropia do solvente. Quando moléculas (ou partes de moléculas) não polares são agrupadas, o tamanho da camada de solvatação é menor, uma vez que nem todas estão expondo toda a sua superfície molecular ao solvente (menor superfície acessível ao solvente). O resultado disso é um aumento favorável na entropia. Consequentemente, aminoácidos hidrofóbicos tendem a se agrupar no interior de uma proteína, mantendo-se afastados da água.

A maior parte da variação da energia livre que ocorre quando as interações intramoleculares são formadas é devido ao aumento da entropia na solução aquosa resultante da formação do núcleo hidrofóbico. Isso supera a grande perda em entropia con-

formacional decorrente do processo de enovelamento da proteína em sua estrutura nativa (Figura 1-7).

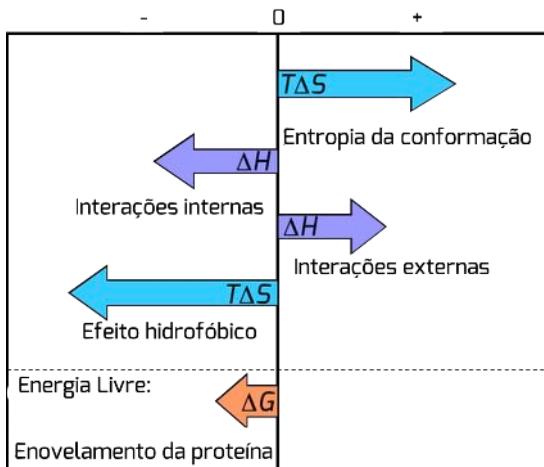



Figura 1-7: A energia livre do enovelamento é resultado de um balanço delicado de forças. As interações intramoleculares (ΔH) e a entropia do solvente (efeito hidrofóbico, TΔS) são favoráveis ao enovelamento, enquanto a entropia conformacional (TΔS) é desfavorável.

7.4. Predição da estrutura

A determinação experimental ainda é considerada o melhor processo para se obter a estrutura tridimensional de uma proteína. Entretanto estas técnicas, além de serem financeiramente custosas, podem levar anos e, em alguns casos, a estrutura final pode não chegar a ser obtida. Portanto, o desenvolvimento de métodos computacionais é tanto uma alternativa mais barata quanto, em alguns casos, a única possibilidade de obtenção de modelos estruturais para algumas proteínas.

A complexidade do estudo das conformações adotadas por uma proteína durante o seu enovelamento até a conformação nativa pode ser ilustrada no chamado de paradoxo de Levinthal. Esse paradoxo diz que o número de possíveis conformações para uma dada sequência de aminoácidos é astronômico exigindo, mesmo considerando os computadores mais poderosos disponíveis, um tempo comparável à idade do universo para o cálculo da energia de todas estas conformações. Entretanto, o tempo de enovelamento de uma proteína está na escala de microssegundos e, portanto, o processo de enovelamento não pode ocorrer através de uma busca aleatória por todas as conformações possíveis. De fato, o que ocorre é a retenção de estruturas que são energeticamente mais estáveis, isto é, a cadeia peptídica percorre um caminho de enovelamento.

Percebe-se, através do paradoxo de Levinthal, porque determinar a estrutura 3D nativa a partir da sequência de aminoácidos permanece como um dos maiores problemas da ciência moderna, tratando-se de uma questão profundamente multidisciplinar e abrangendo diversas áreas da ciência como engenharias, biologia, física, química e computação científica.

Os primeiros métodos desenvolvidos para a predição da estrutura de proteínas eram organizados segundo 3 grupos principais: métodos de modelagem comparativa, de predição de enovelamento (ou *threading*) e predição por primeiros princípios (ou *ab initio*). Essas categorias diferem quanto ao uso das informações disponíveis nos bancos de dados de estruturas tridimensionais de proteínas resolvidas experimentalmente. A modelagem comparativa é a metodologia mais dependente dessas informações, sendo a *ab initio* totalmente independente (Figura 2-7).

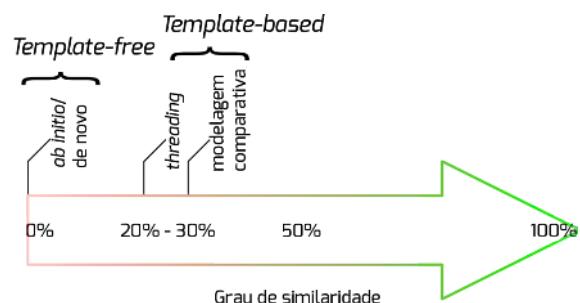



Figura 2-7: Relação entre métodos de predição de estrutura tridimensional de proteínas e o uso de estruturas resolvidas experimentalmente. Cada técnica é aplicável a partir de um certo grau de similaridade, o qual é medido pela taxa de identidade entre os aminoácidos da sequência alvo e sequências de estruturas conhecidas (a serem usadas como moldes).

Com os recentes avanços na área, contudo, pode-se notar que a separação entre entes métodos é cada vez mais tênue. Além disso, uma rápida consulta aos últimos CASP mostra que muitos dos métodos podem ser incluídos em mais de uma categoria. Por exemplo, a separação entre predição do enovelamento e modelagem comparativa é cada vez mais difícil, e o uso de algum tipo de informação estrutural/experimental é amplamente observado, mesmo em metodologias ditas de primeiros princípios. Assim, hoje se usa uma classificação mais ampla que é útil quando se deseja avaliar e comparar os métodos objetivamente:

i) Métodos independentes de estruturas molde (também chamados de métodos *template free*). Incluem a predição *ab initio* e a predição *de novo*;

ii) Métodos baseados em estruturas molde (também chamados de *template based*). Incluem *threading* e modelagem comparativa.

Com esta nova classificação, os métodos ditos *de novo* são aqueles que utilizam algum tipo de informação estrutural, tais como fragmentos de proteínas, predição de estrutura 2ária e potenciais estatísticos, oriundas de proteínas não homólogas à sequência alvo.

O que vai ditar a escolha do método a ser aplicado é a presença ou não de estruturas resolvidas experimentalmente, e depositadas em bancos de estruturas como o PDB (*Protein Data Bank*), que possam ser usadas como molde (ou *template*) para a modelagem da sequência alvo. A escolha do método está intrinsecamente relacionada com a taxa de identidade obtida a partir do alinhamento entre a sequência alvo e possíveis candidatos a molde (Figura 3-7).

O enovelamento da proteína pode ser visto, em última instância, como resultado das forças físicas atuando sobre os átomos da proteína. Sendo assim, a formulação mais acurada para se estudar o enovelamento ou predizer a estrutura de proteína é baseada em representações com todos os átomos explícitos (também chamados de *all-atom*, ver capítulo 8). O problema de tal representação é o nível de complexidade introduzida, que torna o problema muito difícil de ser tratado com a capacidade computacional disponível atualmente. Por razões práticas, a maioria dos métodos de predição faz uso de representações simplificadas da proteína, assim limitando o número de conformações a serem avaliadas (o chamado espaço conformacional), e adotam funções de energia empíricas (ou semi-empíricas) ou baseadas em conhecimento (*knowledge-based*) que capturam as forças mais importantes que impulsionam e estabilizam o enovelamento.

As conformações que estão associadas ao mínimo global da função de energia são consideradas as prováveis conformações nativas que a proteína adota em condições fisiológicas. Dessa forma, os métodos de predição de estrutura de proteínas apresentam, nas suas metodologias, as seguintes características em comum:

i) Uma representação da estrutura da proteína e um conjunto de graus de liberdade que define o espaço de conformações;

ii) Funções de energia compatíveis com a representação;

iii) Algoritmos para realizar a busca no espaço de conformações.

Representação da estrutura e do espaço de conformações

A representação tridimensional de uma molécula pode ser dada pela posição geométrica de seus átomos em um sistema de coordenadas cartesianas (x, y, z) ou pelas chamadas coordenadas internas (Figura 4-7). Nesta última, para cada átomo são fornecidas informações relativas ao comprimento de ligação, ângulo de ligação e ângulo de torção (ou ângulo diedral).

A representação computacional de uma proteína pode ser feita baseada em todos os seus átomos (modelos *all-atom*), em "átomos unidos" (alguns átomos de hidrogênio são considerados implicitamente), e em agrupamentos de átomos (ou *coarse-grained*) (ver capítulo 8). Independentemente da estratégia, as formas de definição são equivalentes.

Sequência Alvo
(formato FASTA)

Busca por Moldes 3D
(alinhamento local)

Identidade > 25% ?

SIM

Modelagem
Comparativa

NÃO

Busca por Tipo de
Enovelamento

Enovelamento
Conhecido ?

SIM

Threading

NÃO

De novo / Ab initio

Modelo 3D

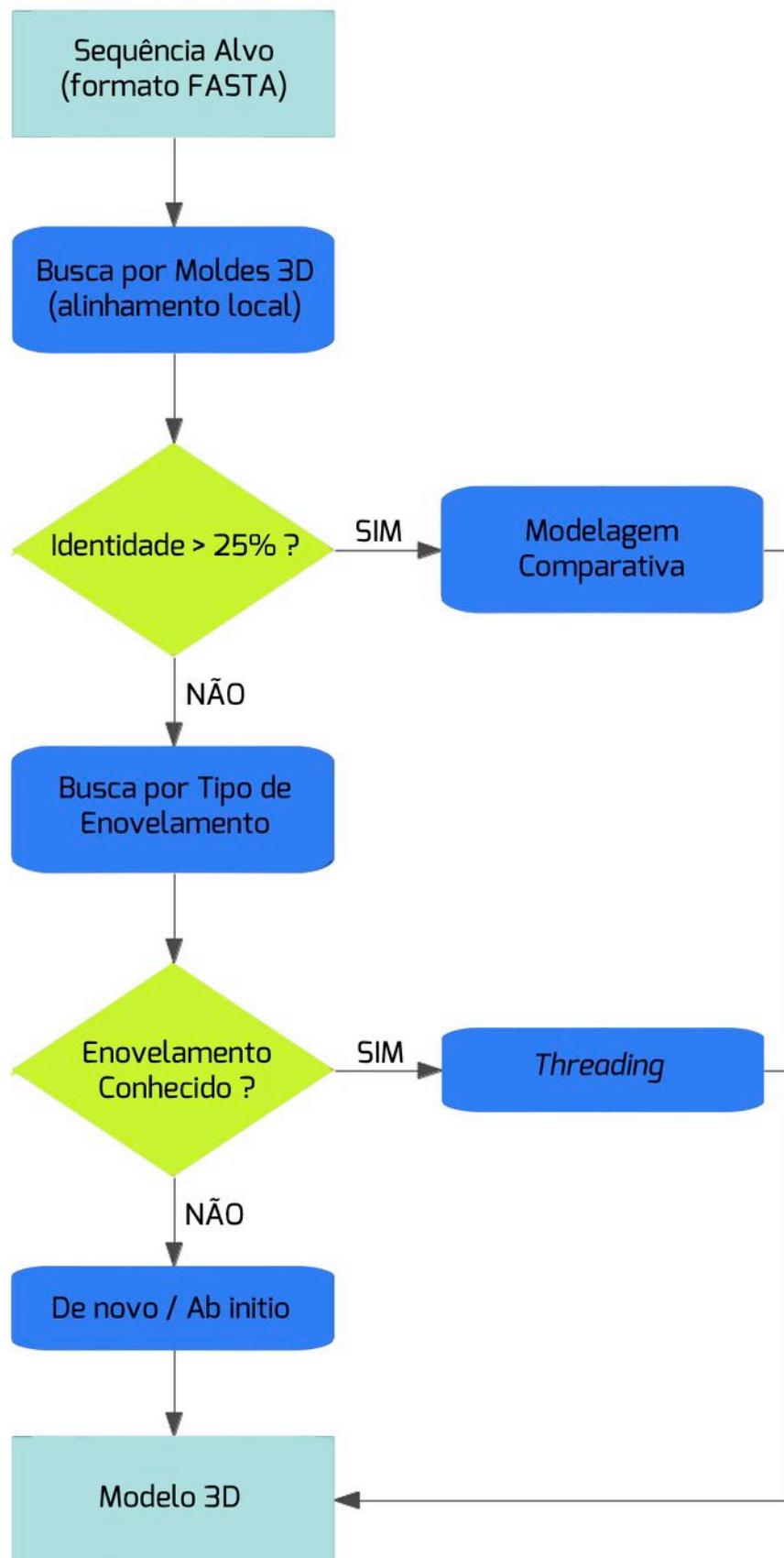



Figura 3-7: Fluxograma para a predição da estrutura tridimensional de uma proteína. O valor de 25% é apenas uma referência e depende de outros fatores, tais como a cobertura com a sequência alvo.

(A) Sistema de coordenadas cartesianas

NOME	NATM	ATM	RES	CAD	NRES	COORDX	COORDY	COORDZ	OCUP	BETA	ELEM
HETATM	1	C1	ETH	A	1	3.108	0.653	-8.526	1.00	0.00	C2H6
HETATM	2	C2	ETH	A	1	4.597	0.674	-8.132	1.00	0.00	C2H6
HETATM	3	1H1	ETH	A	1	2.815	-0.349	-8.761	1.00	0.00	C2H6
HETATM	4	2H1	ETH	A	1	2.517	1.015	-7.711	1.00	0.00	C2H6
HETATM	5	3H1	ETH	A	1	2.956	1.278	-9.381	1.00	0.00	C2H6
HETATM	6	1H2	ETH	A	1	4.748	0.049	-7.277	1.00	0.00	C2H6
HETATM	7	2H2	ETH	A	1	5.187	0.312	-8.947	1.00	0.00	C2H6
HETATM	8	3H2	ETH	A	1	4.890	1.676	-7.897	1.00	0.00	C2H6

(B) Sistema de coordenadas internas

NATM	ATM	BOND	REF1	ANG	REF2	TORC	REF3
1	C						
2	C	1.54	1				
3	H	1.00	1	109.5	2		
4	H	1.00	2	109.5	1	180.0	3
5	H	1.00	1	109.5	2	60.0	4
6	H	1.00	2	109.5	1	-60.0	5
7	H	1.00	1	109.5	2	180.0	6
8	H	1.00	2	109.5	1	60.0	7

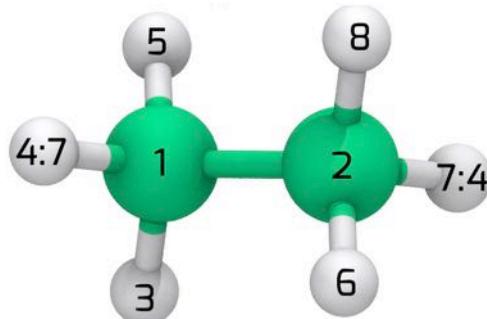



Figura 4-7: Exemplo de representações de uma molécula de etano. Em ambos os sistemas, cada linha representa um átomo. Em A, temos ainda a definição do número de átomos (NATM), do tipo do átomo (ATM), do nome do resíduo (RES), do rótulo da cadeia (CAD), do número do resíduo (NRES) e das coordenadas em si (COORDX, COORDY, COORDZ). Para definição das propriedades descritas em OCUP e BETA, ver capítulo 13. Em B, temos definido o elemento químico (ATM), o comprimento da ligação (BOND), o número do átomo com o qual há a ligação (REF1, por exemplo, o átomo 7 está ligado ao átomo 1, distando deste 1,0 Å), o valor do ângulo de ligação (ANG), o número do átomo com o qual há a formação do ângulo (REF2, por exemplo, o átomo 8 está ligado ao 2 e faz um ângulo de 109,5º com o átomo 1), o valor do ângulo de diedro (TORC) e, por fim, o número do átomo com o qual está definida a torção.

Outro aspecto a ser definido nessa etapa são os graus de liberdade que irão definir o espaço de conformações, isto é, de que forma será definida a flexibilidade estrutural que irá permitir construir diversas estruturas para as sequências alvo. Tipicamente, os métodos de PSP adotam geometrias de ligação rígidas, isto é, o comprimento das ligações é fixo em um valor de referência, assim como os ângulos entre as ligações.

Usando uma representação em coordenadas internas, os graus de liberdade para modificação da estrutura são os ângulos de torção, mais especificamente os ângulos diedrais do esqueleto peptídico: φ, ψ e ω (Figura 5-7, ver também capítulo 2) além dos ângulos diedrais das cadeias laterais: χ_1 até χ_4 (Figura 6-7). A definição desses ângulos é suficiente para construir uma estrutura muito próxima à estrutura nativa de proteínas, de forma muito mais simples do que lidar com o sistema de coordenadas cartesianas.

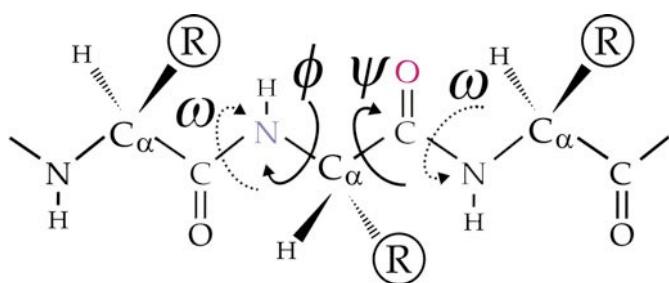



Figura 5-7: Ângulos de torção (diedrais) da cadeia principal da proteína.

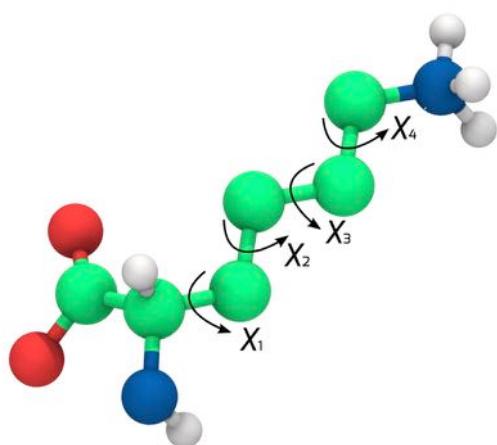



Figura 6-7: Ângulos de torção (diedrais) da cadeia lateral do aminoácido lisina. Até quatro ângulos de torção definem a conformação da cadeia lateral de qualquer aminoácido.

Funções de energia

As conformações geradas pelo algoritmo de predição de estrutura 3D de proteínas devem ser avaliadas seguindo um critério de qualidade. Geralmente, esse critério é dado pela energia total da estrutura. Essa energia pode ser calculada considerando diversos aspectos físico-químicos e diferentes níveis de simplificações. Os parâmetros desta função são usualmente retirados de campos de força clássicos (ver capítulo 8) e, de maneira geral, é uma função dependente da posição dos átomos (ou grupos de átomos) em relação aos seus vizinhos. Nestas funções, a energia total é determinada pela posição dos átomos, e é dada pela combinação das energias fornecidas pelos potenciais diedral próprio, Lennard-Jones e Coulomb (ver capítulo 8).

Algumas abordagens usam funções de energia potencial *ad hoc*, que refletem características gerais das proteínas, e potenciais estatísticos parametrizados a partir de bancos de dados de estruturas conhecidas. Alguns métodos lançam uso de funções efetivas de solvatação que modelam as interações entre a proteína e o solvente (implícito).

De maneira geral, do ponto de vista energético, consideramos a estrutura nativa de uma proteína como sendo a estrutura de menor energia total. Idealmente, a função aplicada deve ser capaz de separar estruturas nativas de não nativas e, além disso, de avaliar o quanto uma estrutura está mais próxima da nativa em relação à outra através da comparação das energias. Tendo em vista esse quadro ideal, a definição da função de energia é um dos aspectos mais difíceis em PSP.

Frequentemente, deve-se decidir entre aumentar a complexidade da função de energia (o que nem sempre garante aumento de precisão) ou usar um modelo mais simplificado para manter um custo computacional que torne o cálculo exequível dentro da infraestrutura computacional disponível. O uso de funções com potenciais estatísticos parametrizados por estruturas conhecidas é uma tentativa de sanar essas dificuldades. No entanto, isso acaba introduzindo outros problemas, como a alta dependência da parametrização e até mesmo a perda de generalidade na aplicação, ou seja, um potencial parametrizado para uma classe de proteínas irá apresentar resultados imprecisos quando aplicado a outra classe.

Algoritmos de busca

O algoritmo de busca é o componente responsável por gerar a conformação inicial, avaliar sua qualidade usando a função de energia, gerar novas conformações e avaliá-las em um processo iterativo até que algum critério de parada esteja satisfeito. O problema de predição de estrutura de proteínas é, geralmente, definido como um problema de minimização. Assim, a busca é feita pela conformação que minimize a função de energia, a qual se espera que seja a conformação nativa.

O problema de otimização possui algumas características que o tornam extremamente complexo. Por exemplo, a função de energia apresenta uma multimodalidade massiva (ou seja, possuem um número muito grande de mínimos locais), degenerescência de mínimos e grandes regiões de conformações inválidas. Além disso, o problema está associado a um número muito grande de graus de liberdade com grande interdependência.

As abordagens empregadas na resolução desse problema frequentemente fazem uso de métodos de-

nominados metaheuristicos (Figura 7-7). Estes métodos constituem-se em técnicas iterativas de otimização nas quais uma solução candidata vai sendo melhorada seguindo uma medida de qualidade. Esses métodos não fazem uso de informações sobre a função de avaliação ou mesmo sobre o problema, no entanto não há garantias de se encontrar a solução ótima. Os métodos metaheurísticos mais comuns incluem aqueles denominados Monte Carlo e Algoritmos Genéticos. No entanto, alguns métodos usam metaheurísticas combinadas a métodos determinísticos baseados no gradiente da função, tais como o método do máximo declive (*steepest descent*). Esses últimos são geralmente aplicados em etapas de refinamento e apenas com funções de energia deriváveis.

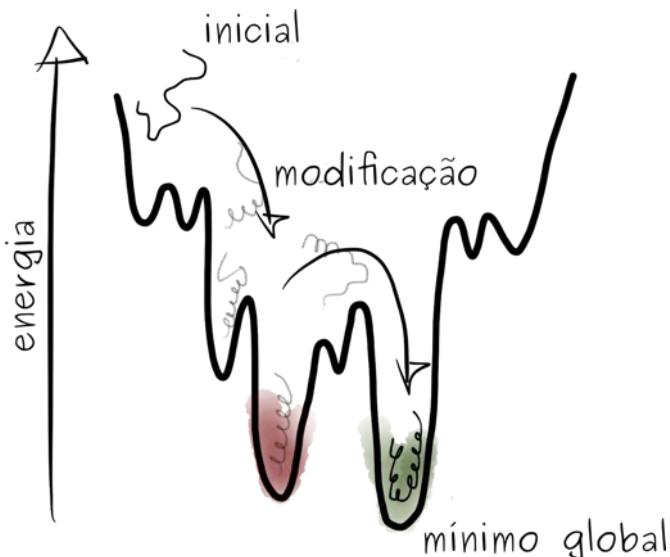



Figura 7-7: Esquema de uma busca usando metaheurística para predição de estrutura de proteína. A estrutura inicial é modificada a cada passo e vai sendo avaliada segundo um critério energético até que se obtenha uma estrutura de mínimo. Idealmente, deseja-se uma estrutura de mínimo global (área em verde) e não uma de mínimo local (área em vermelho).

7.5. Modelagem comparativa

No método de modelagem comparativa, também chamada de modelagem por homologia, a proteína de interesse (alvo) terá sua estrutura 3D predita usando como referência a estrutura 3D de outra proteína similar (também chamada de molde, e na maioria das vezes evolutivamente relacionada). Essa proteína similar tem de possuir estrutura 3D resolvida experimentalmente, e as coordenadas cartesianas de seus átomos devem estar depositadas em banco de dados de estruturas como o PDB.

A modelagem comparativa é o método empregado mais frequentemente, e seu limite de predição está intrinsecamente relacionado com o grau de similaridade entre as estruturas alvo e molde. Geralmente, consideram-se como limites mínimos de aplicabilidade do método valores de 25 a 30% de identidade, obtidos através do alinhamento entre a estrutura 1^{ária} da proteína alvo e de uma ou mais proteínas molde. A modelagem comparativa pode ser dividida em cinco etapas descritas a seguir e resumidas na Figura 8-7.

Identificação de referências

Tem por objetivo identificar sequências de aminoácidos de proteínas resolvidas experimentalmente que possuam similaridade com a sequência da proteína de interesse (sequência alvo), cujas estruturas serão empregadas posteriormente como moldes. Essa identificação pode ser feita através de algoritmos de alinhamento, sendo selecionadas como referências as proteínas que possuírem os maiores índices de similaridade e identidade (suficientes para se inferir homologia entre as sequências), menores índices de *gaps* e a maior cobertura da sequência (relação entre a quantidade de aminoácidos alinhados entre as duas sequências e o tamanho total da sequência alvo).

Seleção dos moldes

Dentre as referências, é necessário escolher uma ou mais estruturas que servirão de molde para a construção do modelo 3D final. Nesta etapa, é imprescindível a análise do papel biológico da proteína de interesse. Os critérios de seleção podem incluir:

i) a proteína de interesse e o possível molde pertencem a uma mesma família de proteínas;

ii) ambas desempenham preferencial-

Sequência Alvo (formato FASTA)

- Alinhamento Local: sequência alvo contra sequências no PDB.
- Avaliar: similaridade, identidade, gaps e cobertura.

Identificação de Referências (busca por moldes 3D)

- Considerar a função biológica.
- Se for enzima, considerar moldes com a presença de ligantes.
- Avaliar qualidade da(s) referência(s).
- Se mais de uma referência, buscar por regiões e resíduos conservados,

Seleção de Molde(s)

- Alinhamento de toda a extensão da sequência alvo com o(s) molde(s).
- Avaliar regiões sem alinhamento: desconsiderar, usar métodos híbridos ou usar restrição de estrutura secundária?

Alinhamento Global (pareado / múltiplo)

- Construção a partir de informações obtidas da(s) estrutura(s) molde(s).
- Técnicas mais aplicadas: corpos rígidos e satisfação de restrições espaciais.

Construção do Modelo

- Avaliar qualidade: estereoquímica, energia, correlação estrutural 1D-3D, hidrofobicidade, conservação de domínio e ligações.

Validação do Modelo

Estrutura OK ?

NÃO

SIM

Modelo 3D

Figura 8-7: Etapas de predição de estrutura tridimensional de proteínas usando o método de Modelagem Comparativa.

mente a mesma função ou tenham funções correlacionadas;

iii) as estruturas resolvidas experimentalmente possuam alta qualidade (por exemplo, resolução ≤ 2 Å, fator R < 20%);

iv) em tratando-se de uma enzima, é recomendado o uso de um molde cuja estrutura já tenha sido resolvida experimentalmente com seu substrato, ligante ou modulador.

Na escolha de mais de uma estrutura molde, é importante realizar o alinhamento estrutural entre estas de forma a identificar regiões conservadas, sítios de ligação, águas estruturais e ligações dissulfeto conservadas.

Alinhamento entre as sequências

Uma vez escolhida(s) a(s) estrutura(s) molde, é necessário realizar alinhamento entre as sequências alvo e molde de forma a garantir que toda a proteína de interesse seja modelada (agora empregando programas como Clustal, T-Coffee e Muscle). Um alinhamento com mais de 40% de identidade é o suficiente para gerar um modelo confiável. Entretanto, é importante lembrar que o modelo final será uma representação desse alinhamento gerado. Portanto, regiões sem alinhamento significativo com o molde são preditas tridimensionalmente (quando preditas) sem grande confiabilidade, usando geralmente dados estatísticos gerais sobre estruturas de proteínas.

Para as regiões sem alinhamento, deve-se considerar:

i) a posição dessa região na sequência de aminoácidos, verificando-se possíveis sítios de clivagem (principalmente em porções N- e C-terminal);

ii) o tamanho dessa porção, considerando-se a possibilidade de formação de um novo domínio até então não identificado nessa família;

iii) se são porções transmembranares, sejam preditas *in silico* (por exemplo, através das ferramentas TMHMM, HMMTOP, TMPred) ou já descritas em literatura porém ausentes nas estruturas molde;

iv) o tipo de estrutura 2ária predita *in silico* por mais de uma ferramenta (tais como PSIPRED, PHYRE, JUFO e PORTER), usando as regiões de consenso entre elas como informação de restrição de tipo de estrutura 2ária durante a etapa de construção do modelo.

Alternativamente, métodos híbridos podem ser aplicados para a predição de porções sem alinhamento. Para essas regiões, aplicam-se os métodos de predição de enovelamento ou primeiros princípios e usa-se a melhor estrutura predita como mais um molde para o método de modelagem comparativa.

Construção do modelo

A partir do alinhamento global entre as sequências alvo e molde, algoritmos específicos para PSP via modelagem comparativa irão transferir as informações extraídas da estrutura 3D da proteína molde para o modelo. As técnicas mais aplicadas são as de construção usando corpos rígidos e por satisfação de restrições espaciais.

A técnica de construção usando corpos rígidos constrói um modelo por partes, baseando-se na conservação de estruturas entre proteínas homólogas ou com grau significativo de identidade. As regiões estruturalmente conservadas da proteína de interesse são definidas através de predição de estruturas 2árias. Essas regiões são alinhadas com o molde, considerando-se a média das posições dos Cα das sequências de aminoácidos das regiões estruturalmente conservadas.

As regiões que não satisfazem as exigências são chamadas de regiões variáveis. Essas compreendem, geralmente, porções de alças que conectam as regiões conservadas. A cadeia principal dessas regiões pode ser obtida em bancos de dados específicos de estruturas, que apresentam conjuntos de alças classificados pelo número de aminoácidos e pelo tipo de estruturas 2árias que conectam.

Após a inserção das regiões de alças, um modelo inicial do esqueleto peptídico estará pronto, restando apenas a inserção das cadeias laterais dos aminoácidos através de busca em bibliotecas de rotâmeros. Como exemplo de programa baseado nesta técnica, pode-se mencionar o portal Swiss-Model.

A segunda técnica mais comum, a construção por satisfação de restrições espaciais, inicia-se pelo alinhamento entre as sequências alvo e molde, extraindo-se desse molde suas restrições espaciais (distâncias e ângulos) e transferindo-as para o modelo. Por exemplo, o tamanho das ligações e seus ângulos preferenciais são obtidos de campos de força. Dessa forma, é possível limitar o número de possíveis conformações que o modelo pode assumir.

A principal característica dessa técnica é a obtenção empírica das restrições espaciais, expressas por funções de probabilidade, a partir de bancos de dados contendo informações sobre alinhamentos entre estruturas proteicas de alta resolução. As restrições espaciais e os termos de energia são combinados em uma função objetivo, sendo submetida a métodos de otimização por gradiente conjugado e recozimento simulado, visando a minimização das violações das restrições espaciais. Como exemplo de emprego desta técnica, pode-se citar o programa Modeller.

Validação do modelo

Após a construção do modelo, é necessário identificar possíveis erros relacionados aos métodos empregados, à escolha das referências e ao alinhamento entre as sequências alvo e molde. Caso o modelo seja caracterizado como de má qualidade, todo o protocolo anterior deve ser revisto no intuito de se melhorar o alinhamento, escolher outros moldes ou até mesmo decidir-se pelo uso de outros métodos. Os principais métodos de validação de um modelo serão descritos adiante (item 7.10).

Por ser dependente de uma estrutura 3D resolvida experimentalmente, a técnica de modelagem comparativa possui certas limitações, tais como:

i) nem sempre se consegue uma estrutura molde para a proteína de interesse;

ii) o grau de similaridade conseguido entre as sequências alvo e molde pode ser pequeno (<30% de identidade), mesmo em regiões do sítio ativo, inviabilizando o emprego desta técnica;

iii) por vezes, as sequências que podem servir como moldes possuem qualidade insuficiente para a construção de um modelo adequado.

Nesses casos, como citado anteriormente, o uso adicional de informações, como a identificação de regiões transmembranares, a predição de regiões de peptídeo sinal, a predição de tipo de estrutura 2ária, a predição do tipo de enovelamento e a verificação da existência de dados teóricos e experimentais quanto à existência, quantidade e localização de porções transmembranares, ligantes e número e tipo de cadeias podem contribuir tanto na construção de modelos tridimensionais como na anotação funcional de sequências.

No caso de análises em larga escala de conjuntos de proteínas, e até mesmo de genomas inteiros, todo esse processo deve ser realizado para cada proteína de interesse. Considerando o tempo gasto em cada uma dessas etapas, é interessante o uso de métodos automatizados que podem ser empregados como um filtro inicial para a detecção de quais proteínas podem ser modeladas por modelagem comparativa e para a obtenção de um modelo inicial para cada uma dessas proteínas, a ser otimizado individualmente. Como exemplo de programa usado para a análise em larga escala de sequências de proteínas, citamos o programa MHOLline.

7.6. Predição do enovelamento

O método de predição do enovelamento ou *threading* parte da ideia de observações de que a estrutura 3D é mais conservada que a sequência, de forma que mesmo sequências com pouca similaridade podem possuir estruturas muito semelhantes, o que limita o número de enovelamentos que proteínas podem assumir. Atualmente, mais de 1.000 tipos de enovelamento já foram registrados, e acredita-se que esse valor não ultrapasse a previsão máxima de 7.000 tipos.

Nesse método, também são usadas proteínas com estruturas 3D conhecidas e depositadas no PDB, de onde as informações sobre os tipos de enovelamento são extraídas e armazenadas em bancos de dados de tipos de enovelamentos. Como exemplo, citamos o CATH (*Class*, *Architecture*, *Topology*,

Homology) e o SCOP (*Structural Classification of Proteins*).

O método de predição do enovelamento é assim menos dependente da proximidade evolutiva entre a sequência de aminoácidos da proteína de interesse e seus possíveis moldes, ou seja, as sequências podem apresentar baixa identidade. O método é portanto aplicável quando o alinhamento entre a estrutura $1^{ária}$ da proteína de interesse e de uma ou mais proteínas de referência (moldes) apresentam uma identidade entre 20% e 30%.

No problema de PSP via predição do enovelamento tenta-se ajustar a estrutura $1^{ária}$ da proteína de interesse aos tipos de enovelamentos de proteínas conhecidos, analisando principalmente as conservações de estruturas $2^{árias}$. Esse método pode ser dividido nas seguintes etapas:

i) Reconhecimento do tipo de enovelamento pela análise das principais propriedades da proteína de interesse (tais como estrutura $2^{ária}$, polaridade de cadeias laterais e hidrofobicidade);

ii) Construção do melhor alinhamento possível entre a sequência de aminoácidos da proteína de interesse e estruturas depositadas em bancos de dados. Alguns métodos baseiam-se na construção de modelos simplificados (como modelos baseados em Cα) da proteína de interesse a partir da estrutura 3D de possíveis moldes, e avaliam a qualidade do modelo através da otimização de funções objetivo (geralmente não-lineares). Essas funções podem considerar, por exemplo, resultados de alinhamentos múltiplos de sequências e de estruturas $2^{árias}$, matrizes de substituição para cada aminoácido dentro de uma família específica de proteínas e penalização de *gaps*;

iii) Escolha do(s) melhor(es) molde(s) para a construção da estrutura 3D da proteína de interesse, geralmente baseada em funções de predição de erro/qualidade entre os possíveis modelos simplificados e seu(s) molde(s) (por exemplo, a função TM-score). A escolha dos melhores moldes por vezes é baseada em bibliotecas de fragmentos;

iv) Construção do modelo 3D através de técnicas similares às empregadas na modelagem comparativa, por vezes valendo-se de ferramentas acopladas aos programas Swiss-Model ou Modeller. Alguns programas empregam, para as regiões sem molde, métodos por primeiros princípios. Como exemplo de programas para PSP via predição do enovelamento pode-se citar os programas HH-Pred e I-TASSER.

As limitações dos métodos de predição do enovelamento vêm de dois pontos principais. O primeiro é similar ao observado para a modelagem comparativa, isto é, se a identidade entre a sequência alvo e as proteínas utilizadas na construção do banco de enovelamentos for muito baixa, é possível que o enovelamento daquela sequência simplesmente não esteja representado no banco. Assim, o método pode construir um modelo completamente errado. A outra limitação é que os modelos apresentam uma resolução relativamente baixa, dificultando seu uso em estudos que exigem posicionamento preciso dos átomos como no caso do atracamento (ver capítulo 9).

7.7. Métodos *de novo*

Embora a modelagem comparativa e a predição do enovelamento permitam a obtenção de modelos satisfatórios, tais técnicas são inválidas se proteínas de referência, com estruturas determinadas experimentalmente, não se encontrarem disponíveis. De forma a manter a independência de moldes de proteínas homólogas, foram desenvolvidos métodos que usam informações provenientes de bancos de estruturas de proteínas determinadas empiricamente, sem a necessidade de haver identidade com a sequência alvo, resultando na predição chamada *de novo*. Dentre as principais técnicas usadas pela predição *de novo* destacam-se o uso da predição de estruturas $2^{árias}$, uso de fragmentos de proteínas, e modificação da função de energia.

Predição de estruturas $2^{árias}$

A predição de estruturas $2^{árias}$ envolve o conjunto de técnicas que visam reconhecer as categorias de estruturas 2^{as} (tipicamente hélices e folhas) associadas a cada região de

uma proteína a partir apenas de sua sequência. Por 30 anos, o cenário de técnicas de predição de estruturas 2^{árias} foi composto por métodos que se baseavam na propensão de um resíduo pertencer a uma determinada estrutura 2^{ária}. Na década de 1990, uma nova geração de métodos que considerava os efeitos trazidos pelos resíduos adjacentes surgiu, contemplando os efeitos de interações locais na predição, o que alçou a precisão das predições a um patamar acima de 60%.

O crescimento de bancos de dados de proteínas em combinação a algoritmos mais sofisticados permitiu a inclusão de informações relacionadas ao enovelamento da proteína nestas predições, principalmente aquelas relacionadas aos efeitos de interações de longo alcance. Esses novos métodos baseiam-se em alinhamentos múltiplos e sua consequente informação evolutiva. Em sua maioria, esses métodos valem-se do PSI-BLAST (ver capítulo 3). Os atuais métodos de predição de estruturas 2^{árias} possuem desempenho em torno de 80% de precisão, dentre os quais destacam-se PSIPRED, DSC, GOR IV, Predator, Prof, PROFphd e SSpro.

Fragmentos de proteínas

A determinação da estrutura da RBP (Retinol Binding Protein) em 1986, em particular de seu sítio ativo, se mostrou desafiadora por sua estrutura não se parecer com nenhuma até então conhecida (Figura 9-7). Diante das dificuldades de se concluir tal trabalho de determinação, os pesquisadores resolveram buscar informações em todo o banco do PDB (na época contava com apenas 213 entradas), procurando por quaisquer estruturas (ou regiões/segmentos destas) semelhantes que pudesse substituir o sítio em estudo. Nessa busca, os autores perceberam que a segmentação das proteínas em pequenos fragmentos resultava em uma surpreendente redundância estrutural, ou seja, pequenos fragmentos com estruturas similares apresentavam similaridade de sequência (localmente).

Isso permitiu a construção de um mo-

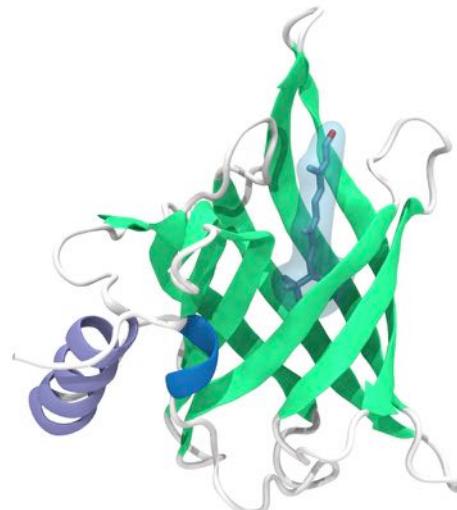



Figura 9-7: *Retinol Binding Protein* com o retinol no sítio ativo, código PDB: 1RBP.

delo da RBP a partir de fragmentos de outras proteínas, sem qualquer grau de similaridade global, e previu-se que se tratava de uma proteína organizada em uma série de oito fitas β antiparalelas, constituindo um barril-β que encapsula a molécula de retinol. A facilidade com que uma estrutura, então considerada incomum, foi prevista usando-se estruturas parciais de muitas proteínas diferentes levou os autores a questionarem se haveria alguma estrutura de proteína que pudesse de fato ser considerada única, e a proposta desta técnica de modelagem por meio de fragmentos proteicos cujas estruturas tivessem sido determinadas experimentalmente (ou seja, empiricamente).

A preservação de certo grau de similaridade estrutural entre trechos curtos de sequências semelhantes é a chave para a predição na ausência de moldes (*template-free*) de estruturas de proteínas. Quando não há qualquer proteína homóloga disponível para ser usada como molde, é possível usar um conjunto de pequenos fragmentos que se correlacione localmente com a estrutura da proteína alvo (Figura 10-7).

Entretanto, deve-se perceber que por maior que seja a similaridade entre duas sequências de fragmentos, a similaridade estrutural é apenas parcial. Como cada fragmento, sendo proveniente de uma proteína diferente, encontra-se imerso em um ambiente físico-químico próprio, o conjunto de

interações que agem sobre esses fragmentos podem conferir-lhes estruturas diferentes (Figura 11-7).

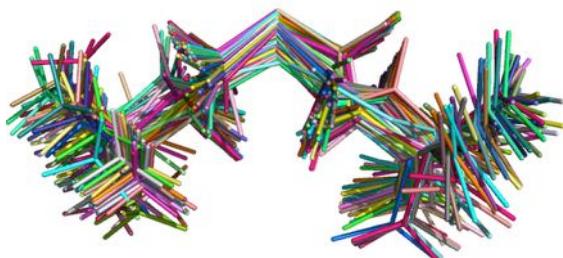



Figura 10-7: Fragmentos estruturalmente semelhantes, mas que possuem sequências de resíduos diferentes.

Duas características devem ser levadas em consideração para se trabalhar com fragmentos de proteínas na predição de estruturas: a primeira é que uma mesma sequência pode levar a estruturas diferentes, e a segunda é que duas sequências diferentes podem levar à mesma estrutura. Dessa forma, se faz necessário a construção de uma lista de fragmentos candidatos a reproduzir uma dada região da proteína alvo.

O primeiro desafio para a predição de estruturas usando fragmentos é montar uma biblioteca de fragmentos que reúna as melhores estruturas candidatas a reproduzir a região da sequência alvo, a partir de um banco de proteínas determinadas empiricamente. Como discutido anteriormente, pode-se usar a similaridade entre as sequências dos fragmentos retirados das proteínas do banco e a região de interesse da proteína alvo. Os programas Rosetta e QUARK usam o PSI-BLAST para reconhecer o quão similares são as sequências de um fragmento e da respectiva região da proteína.

Como exemplo da geração de uma biblioteca de fragmentos podemos citar o programa *Protein Fragment Generator* - Profrager. Nele, os fragmentos são extraídos de uma versão do PDB filtrada para eliminar as diversas redundâncias existentes entre as estruturas. Cada fragmento é iniciado em um resíduo da proteína e se estende pelo comprimento desejado. Uma biblioteca de fragmentos, por exemplo de 6 resíduos, compreende os resíduos das posições 1 a 6, 2 a 7, 3 a 8 e assim sucessivamente. De posse dos frag-

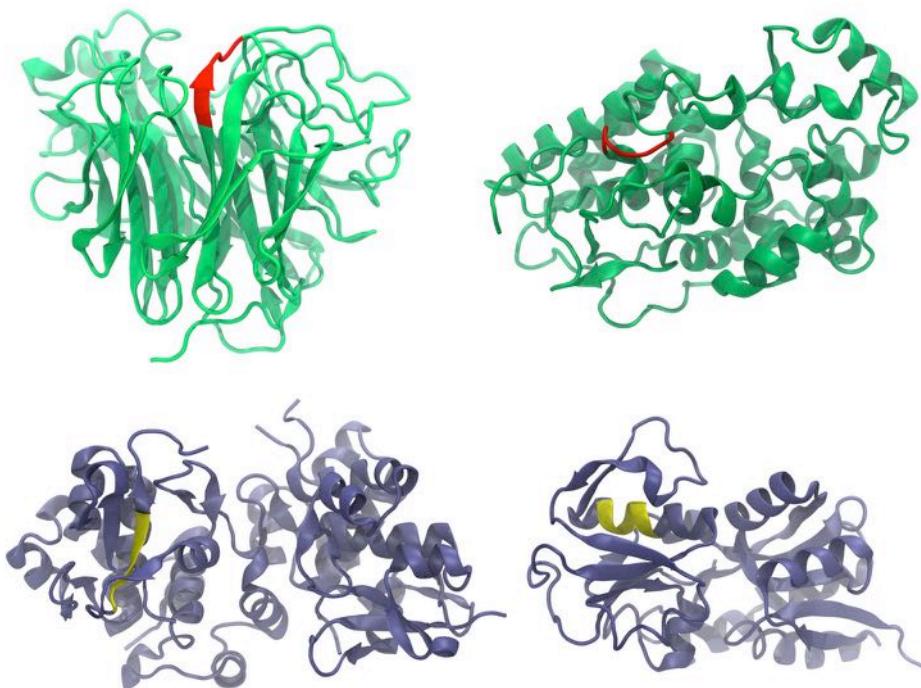



Figura 11-7: Fragmentos de proteínas com a mesma sequência de resíduos que possuem estruturas diferentes. Acima, as proteínas de código PDB 1F8E (fragmento destacado entre os resíduos 243 e 247) e 1BGP (resíduos 63 a 67); abaixo, 1LM5 (2800 a 2804) e 1XS5 (121 a 125).

mentos extraídos do banco, o problema torna-se então escolher os melhores para reproduzir cada região.

Na Figura 12-7 está representada uma biblioteca com fragmentos de 6 resíduos para uma dada proteína. O primeiro fragmento do banco é alinhado à primeira posição da proteína. Os resíduos do fragmento são comparados com as entradas da matriz BLOSUM62. Nesse exemplo, o valor da substituição de uma valina por uma asparagina é -3, e a substituição de um glutamato por uma lisina é +1. Somando os valores da comparação entre todos os resíduos do fragmento com os da respectiva região da sequência alvo, temos uma pontuação total de -8 para esse fragmento. O segundo fragmento do banco é tomado, e o processo de comparação resíduo-resíduo entre o fragmento e a sequência alvo é repetido. Nesse exemplo, tem-se uma pontuação total de +11 para o segundo fragmento. O processo ilustrado para a atribuição da pontuação é repetido para todos os fragmentos do banco, sempre para uma janela de leitura de 6 resíduos. Ou seja, desloca-se um resíduo para a direita e reinicia-se o processo, formando uma nova lista de fragmentos para esta nova posição.

Uma lista de candidatos a reproduzir a sequência alvo é montada de acordo com uma pontuação. Parte dessa pontuação é o grau de similaridade entre a sequência do fragmento e da região correspondente da sequência alvo. A outra parte da pontuação é a concordância da estrutura $2^{ária}$ do fragmento com a estrutura $2^{ária}$ predita pelo PSIPRED para a sequência alvo. Ao final, a biblioteca de fragmentos conterá os fragmentos que possuem as maiores pontuações, logo, os fragmentos mais prováveis para a reprodução da estrutura local.

Se o uso de um fragmento de uma proteína conhecida elimina a necessidade de se modelar a região localmente, o problema torna-se escolher a melhor estrutura para cada região. De posse de uma biblioteca de fragmentos, o trabalho torna-se um problema de otimização, abordado por um algoritmo de busca, onde se procura reconstruir a proteína usando as informações trazidas pelos fragmentos, validando-se a estrutura gerada usando uma determinada função de energia.

É importante notar que, embora sejam dependentes de bancos de estruturas, os fragmentos não precisam ser provenientes de proteínas com grau elevado de identidade, o que permite a modelagem de estruturas inéditas. Modelos obtidos com o uso de fragmentos demonstram utilidade para inspirações biológicas e têm obtido sucesso nas demais áreas da modelagem de proteínas, tais como predição de sítios ativos e identificação de padrões de enovelamento, atracamento proteína-proteína, modelagem de voltas e até mesmo desenho de novas proteínas.

As limitações dos métodos *de novo* são praticamente as mesmas dos métodos por primeiros princípios. Sua aplicação é, em geral, limitada a sequências mais curtas (<150 resíduos), e alguns dos métodos podem estar sujeitos a artefatos se a parametrização das funções estatísticas não for feita com cuidado.

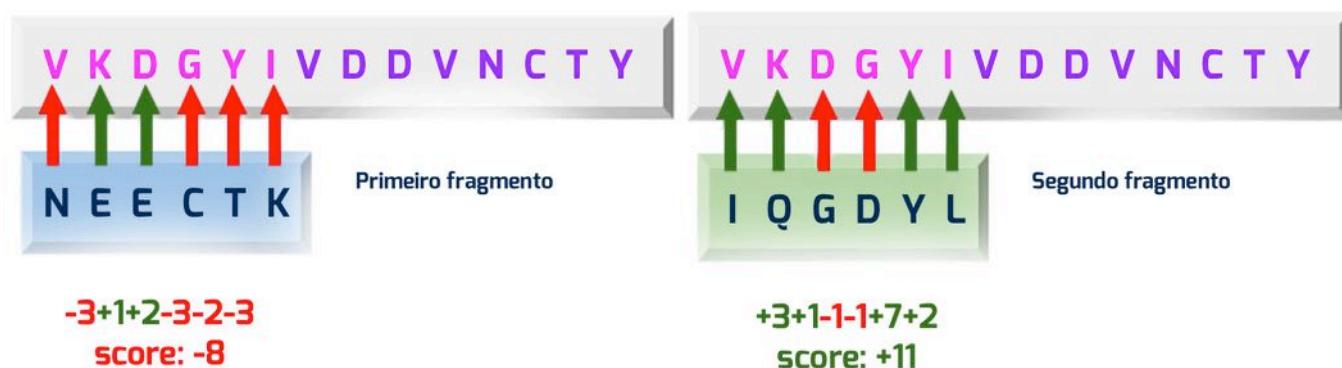



Figura 12-7: Geração de um fragmento de seis resíduos.

Campos de força estatísticos

Campos de força clássicos (ver capítulo 8) são comumente empregados para a representação de interações intramoleculares da estrutura de proteínas, como ângulos e comprimentos de ligação, ângulos diedrais, forças de van der Waals e eletrostáticas. Entretanto, os métodos de maior sucesso nos últimos anos para predição da estrutura de proteínas empregam termos estatísticos derivados de proteínas cujas estruturas já são conhecidas, seja de forma exclusiva ou combinados com termos de campos de força clássicos. Isso culmina nos chamados campos de forças estatísticos, cujo desenvolvimento se tornou amplamente disseminado.

Uma das formas de representar o universo de conformações que uma determinada sequência polipeptídica pode adotar é através de uma superfície, onde cada ponto representa uma dada conformação. Nesta superfície, a altura de cada ponto representa a energia da conformação, de forma que conformações de menor energia estarão no fundo da superfície, e conformações de maior energia em seu topo.

Assim, os termos de campos de força estatísticos são derivados usando-se um conjunto de proteínas teste com a intenção de suavizar a superfície de energia, garantindo que a conformação de menor energia (ou mínimo global) corresponda à conformação nativa, e os mínimos locais sejam pouco frequentes e com valores de energia distantes do mínimo global. A configuração ideal de uma função de energia faz com que as barreiras entre os mínimos sejam menores, permitindo ao algoritmo de busca a passagem de um mínimo local a outro, facilitando a busca pelo mínimo global (Figura 13-7).

Tomemos como exemplo um dos termos mais comuns nas funções de energia, as ligações de hidrogênio. Alguns autores descreveram que é possível gerar todas as estruturas contidas no PDB a partir de um conjunto de representações de ligações de hidrogênio, o que torna interessante um termo do campo de força exclusivamente dedicado ao tratamento dessas ligações. Já se verificou que os termos usados em campos de força clássicos não são capazes de representar todas as ligações de hidrogênio em suas orientações corretas. Assim, um termo estatístico exclusivo para ligações de hidrogênio se mostra fundamental para a predição *de novo*.

Figura 13-7: Efeito de suavização da superfície de energia.

Este termo contribui na avaliação da propensão de formação de estruturas 2^árias^ (ver capítulo 2), usando o valor da probabilidade de um par de resíduos $P(a_i a_j)$ possuir uma ligação de hidrogênio. A probabilidade pode ser calculada de acordo com a equação abaixo:

$$P(a_i a_j) = -log[F_0(a_i a_j)/F_e(a_i a_j)]$$

onde $a_i a_j$ é o par de resíduos, $F_0(a_i a_j)$ é a frequência observada para as ligações de hidrogênio entre os resíduos avaliados e $F_e(a_i a_j)$ é a frequência estimada a partir de um conjunto de estruturas enoveladas incorretamente.

Um exemplo de aplicação destes termos estatísticos é o programa QUARK, um dos métodos de maior sucesso no CASP. É relatado que sua capacidade de refinar estruturas é devida à parametrização de seu campo de força, puramente estatístico. A correlação entre a energia e a similaridade estrutural com a conformação nativa segundo o QUARK é de 0,7 (sendo 0,0 a pior correlação possível e 1,0 uma correlação perfeita).

7.8. Primeiros princípios

A predição por primeiros princípios ou

ab initio se destaca como sendo a tentativa mais ambiciosa para a resolução do problema de predição de estrutura de proteínas. Essa abordagem difere-se das demais por não usar informações de estruturas conhecidas, relacionadas com a sequência alvo, e por usar funções de energia contendo somente termos de significado físico. Tal estratégia é baseada em dois pressupostos: todas as informações necessárias sobre a estrutura de uma proteína estão contidas em sua sequência de aminoácidos, e acredita-se que as proteínas enovelam-se para um estado nativo, ou um conjunto de estados nativos, que se encontra no (ou próximo ao) mínimo global de energia livre.

Além de prever a estrutura tridimensional, os métodos por primeiros princípios podem contribuir na compreensão dos princípios físicos do processo de enovelamento. Adicionalmente, podem ser aplicados na correção ou refinamento de estruturas modeladas por outras metodologias ou mesmo na predição de proteínas desordenadas. O sucesso dos métodos destas predições depende, principalmente, de uma função de energia acurada, na qual o estado nativo da proteína corresponda ao estado termodinamicamente mais estável, e de um algoritmo eficiente capaz de varrer a superfície de energia (ou seja, gerar diversas novas conformações).

O enovelamento de uma proteína pode ser visto, em última instância, como resultado das forças físicas atuando sobre os átomos da proteína. O campo de força deve capturar, ao menos, informações qualitativas essenciais das características físicas e químicas que impulsionam e estabilizam o enovelamento, descrevendo as interações intramoleculares da proteína e desta com as moléculas de solvente. Normalmente, usam-se campos de força empíricos, muitas vezes complementados por um termo de solvatação implícita (ver capítulo 8). Tais funções de energia invariavelmente sofrem aproximações que resultam em artefatos nos modelos, tais como o favorecimento excessivo de estruturas 2^{árias} em hélices em relação a outros tipos de estruturas.

Apesar do alto grau de complexidade, a formulação mais realista para se estudar o enovelamento ou prever a estrutura de proteínas seria baseada em representações com todos os átomos explícitos (ver capítulo 8). Contudo, a predição por primeiros princípios implica em altíssimo custo computacional, e o número de conformações possíveis para uma sequência de aminoácidos é muito grande para ser exaustivamente amostrado. Por isso, parte destes métodos faz uso de modelos de energia e representações simplificadas, tais como modelos *coarse-grained* (ver capítulo 8), acelerando a busca conformacional.

Os algoritmos de busca mais usados são aqueles que envolvem abordagens heurísticas, com destaque para os algoritmos genéticos. Há, também, estudos de predição por primeiros princípios envolvendo o uso de simulações por dinâmica molecular (ver capítulo 8) com campos de força clássicos, apesar de essa técnica ser mais aplicada a estudos do enovelamento proteico.

O primeiro marco na tentativa da predição por primeiros princípios através de simulações por dinâmica molecular foi, provavelmente, nos trabalhos de Duan e Kollman, em 1998, com a simulação da proteína *villin headpiece* (36 resíduos) em solvente explícito, a qual envolveu seis meses de computação paralela em larga escala (projeto *Folding@home*).

A predição por primeiros princípios ainda é um problema não resolvido na biologia computacional. Ela representa a abordagem mais complexa e difícil dentre os métodos de predição e ainda está defasada, em termos de velocidade e acurácia, quando comparada com os demais métodos. Atualmente, seu sucesso é limitado a proteínas pequenas, com menos de 100 resíduos de aminoácidos (Figura 14-7).

Uma grande variedade de métodos vem sendo proposta com dois focos importantes: rapidez e acurácia. A maioria busca o equilíbrio entre esses dois fatores. As diferenças entre esses métodos (Tabela 1-7) se encontram no tipo de representação (ou seja, todos os átomos ou modelos *coarse-grained*), no método de busca e na função de energia.

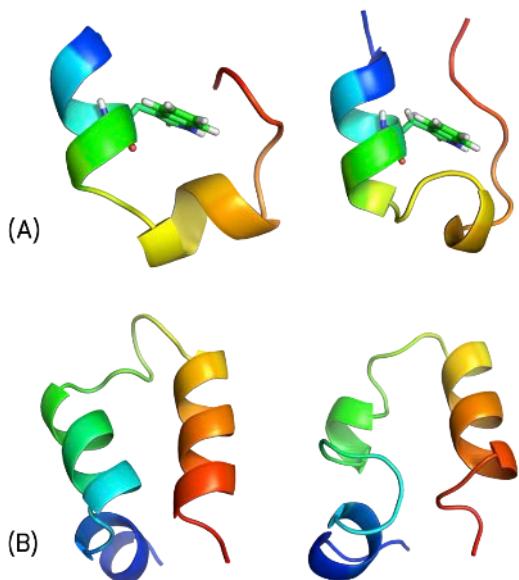



Figura 14-7: Resultados obtidos com o protocolo *ab initio* do programa GAPF. (A) Trp-*cage* (PDB1L2Y) com 29 aminoácidos. O modelo na esquerda apresenta um RMSD (do esqueleto peptídico) de 3,04 Å em relação à estrutura experimental na direita. (B) *Villin headpiece* (PDB1VII) com 36 aminoácidos. O modelo na esquerda apresenta um RMSD de 3,38 Å (do esqueleto peptídico) em relação à estrutura experimental na direita.

Tabela 1-7: Exemplo de métodos de predição *ab initio* de estrutura de proteínas.

Método	Algoritmo de busca	Função de energia
GAPF	Algoritmo genético	GROMOS96 e GAPF-CG
Profet	Algoritmo evolucionário	OPLSAA, AMBER94, AMBER96, ECEPP e FLEX
ProtPred	Algoritmo evolucionário	CHARMM (v.27)
Nicosia	Algoritmo evolucionário multiobjetivo	CHARMM (v.27)
MEAMT	Algoritmo evolucionário multiobjetivo multitabelas	CHARMM (v.27)

7.9. Escolhendo o modelo

Tanto os métodos *de novo* (baseados ou não em moldes) quanto os por primeiros princípios têm em comum a grande quantidade de modelos gerados. Devido à natureza estocástica dos algoritmos de busca (e também às imprecisões das funções de energia), os procotolos mais usados em PSP consistem em executar o algoritmo um grande número de vezes com diferentes sementes para o gerador de números aleatórios. Para efeitos de ilustração, um protocolo típico considerado próprio para publicação do método Rosetta consiste em, no mínimo, 10.000 execuções independentes. Dessa forma, cada execução irá percorrer uma trajetória diferente no espaço de conformações e poderá terminar em uma conformação diferente.

As estruturas resultantes dessa grande amostragem são chamadas de *decoys*, e um problema em aberto na PSP é a filtragem de *decoys*. Atualmente, os protocolos seguem alguns passos para a escolha do modelo a ser selecionado dentre as milhares de conformações geradas. Os dois principais passos são:

i) Filtragem dos *decoys*: é feita sobre o valor de energia total retornado pela função usada pelo método.

A maioria dessas funções já carrega de forma implícita (ou explícita) medidas sobre a qualidade estereoquímica da estrutura. Dessa forma, um primeiro filtro razoável é investigar apenas os *decoys* com energia semelhante (até certo valor limite) em relação ao *decoy* de menor energia (o melhor segundo o critério energético).

ii) Agrupamento (*clustering*) dos *decoys*: é a comparação entre as estruturas resultantes do passo anterior e o seu agrupamento de acordo com um critério de similaridade, por exemplo, estruturas com até 3 Å de RMSD são colocadas em um mesmo grupo. Assim, o pesquisador pode investigar apenas a estrutura mais representativa de cada grupo.

Esse passo tem o potencial de reduzir consideravelmente o número de modelos a ser investigado, embora em alguns casos o número de conformações a

ser analisado possa ainda ser grande demais. Nesses casos, faz-se uso da noção de que, sendo o estado nativo cineticamente acessível, espera-se que esse seja atingido com mais frequência, salvo em trajetórias que terminem em mínimos locais muito profundos. Sendo assim, realizando um número grande de trajetórias, aquele grupo que contém a estrutura nativa é, provavelmente, o maior grupo (ou seja, aquele que contém o maior número de conformações após o agrupamento). É importante ressaltar que esta hipótese só estaria teoricamente bem fundamentada caso usasse uma função de energia realística e representativa da energética do processo de enovelamento.

Os pacotes de PSP disponibilizam suas próprias ferramentas de agrupamento. Pode-se, ainda, usar outros programas externos com resultados semelhantes, como o maxcluster e o programa de agrupamento contido no pacote GROMACS (g_cluster).

Um terceiro passo é a inspeção manual por um operador humano de cada modelo resultante do segundo passo. Com a análise de especialistas treinados, é possível detectar possíveis erros no enovelamento e até mesmo sugerir modificações em regiões específicas dos modelos. Essa etapa opcional ainda não é automatizável sendo, de certa forma, a mais custosa.

7.10. Análise de qualidade

A qualidade de um modelo é determinada por um conjunto de fatores, tais como comprimentos de ligação, planaridade das ligações peptídicas, planaridade dos anéis e ângulos de torção nas cadeias principal (ou seja, esqueleto peptídico) e laterais, quiralidade, impedimento estérico, energia e funcional. Adicionalmente, nos métodos baseados no uso de estruturas moldes resolvidas experimentalmente, para um modelo ser considerado de boa qualidade é recomendado que o valor de RMSD obtido pela sobreposição da cadeia peptídica de regiões conservadas do modelo gerado e da estrutura molde esteja entre 1 Å e 2 Å. Dentre as análises a serem feitas, recomenda-se as seguintes:

i) Estereoquímica: consiste em analisar os aspectos tridimensionais de uma molécula, a fim de se verificar a estabilidade conformacional da mesma. Nesta análise, são detectadas regiões de tensão angular e torcional, impedimentos estéricos e quiralidades. Além destes, com a análise do gráfico de Ramachandran é possível identificar, através da correlação entre os ângulos φ e ψ, quais resíduos encontram-se fora das regiões energeticamente favoráveis, possibilitando uma melhora no modelo final. Exemplos de programas que realizam estas análises incluem os programas Procheck e Molprobity.

ii) Energia: são métodos baseados em minimização de funções de energia. A análise dos valores normalizados da função (como o DOPE normalizado do Modeller) ajuda a avaliar (ao menos estatisticamente) quão próximo o modelo gerado está de proteínas que possuem um mesmo perfil molecular ou até o mesmo tipo de enovelamento. Esses métodos podem considerar a relação entre a estrutura 1D-3D, ponderar a propensão de cada aminoácido estar em um tipo de estrutura 2^{ária}, a probabilidade de dois resíduos estarem em contato e até mesmo o tipo de função que a proteína desempenha. Alguns programas bastante usados para estas análises incluem Verify3D, ProSa, QMEAN e PROVE.

iii) Funcional: envolve a comparação do modelo obtido com aspectos funcionais ou mesmo estruturais (sem resolução atomística) determinados por métodos experimentais. Por exemplo, diversas famílias de proteínas possuem resíduos específicos associados à função (como a tríade catalítica em serino proteases ou resíduos ligadores de metais em metaloproteínas). Assim, o modelo gerado deve apresentar tais resíduos nas suas localizações específicas para explicar dados experimentais prévios. Ainda, métodos como dicroismo circular (capítulo 10), infravermelho (capítulo 11) e

RMN (capítulo 12) podem oferecer informações importantes sobre o estado conformacional da proteína em meio biológico, validando o modelo obtido. Mesmo que as estratégias de análise anteriores indiquem um modelo de elevada qualidade, se o mesmo não for capaz de apresentar ou explicar características conhecidas previamente, não poderá ser considerado totalmente válido.

Durante o CASP a análise de qualidade dos modelos assume um caráter diferente, uma vez que os avaliadores conhecem a estrutura nativa. Nesse caso, a métrica empregada para comparar a estrutura nativa com os modelos gerados pelos diferentes métodos é o *Global Distance Test* – GDT. Trata-se de uma medida potencialmente mais acurada, uma vez que é menos sensível a discrepâncias muito grandes, oriundas de regiões de voltas que são naturalmente flexíveis.

7.11. Refinamento do modelo

Após a análise do modelo, caso a qualidade não tenha sido satisfatória, algumas estratégias de refinamento no melhor modelo obtido podem ser suficientes para a obtenção de um modelo final de boa qualidade. Dentre os principais tipos de refinamento podemos citar:

i) Local: através da análise estereoquímica pode-se identificar qual resíduo está violando seus valores limites dentro de sua vizinhança, o que geralmente é resolvido com o reposicionamento de sua cadeia lateral. Em alguns casos, é necessário realizar etapas de otimização somente de regiões de alças, principalmente de regiões ricas em glicina. É sempre importante observar violações causadas por prolinas nas extremidades de regiões de estruturas em hélice ou folha.

ii) Imposição de restrições: após a análise de resultados de métodos de predição de estrutura 2ária, pode-se verificar no modelo gerado quais regiões não possuem ou possuem uma baixa similaridade de sequência com o(s) molde(s) usado(s), ou não obedecem ao tipo correto de estrutura 2ária predita. Para corrigir isso, é necessário refazer o modelo 3D impondo ao algoritmo de construção o uso de restrições de tipo de estrutura 2ária para essas regiões.

iii) Dinâmica molecular: Os métodos de simulação por dinâmica molecular (ver capítulo 8) têm sido empregados na melhora de modelos gerados tanto por técnicas baseadas em modelagem comparativa quanto por primeiros princípios. Simulações em solvente explícito ajudam a acomodar a estrutura 3D do modelo melhorando, principalmente, os ângulos φ e ψ de resíduos em regiões desfavoráveis no gráfico de Ramachandran. O tempo de simulação é variável de acordo com a complexidade do sistema e com o grau de refinamento que se deseja obter. É importante destacar que simulações por dinâmica molecular para estruturas transmembranares, apesar de bastante recomendado, necessitam especial atenção, pois se deve considerar o modelo de membrana a ser empregado, a forma de inserção do modelo 3D da proteína na membrana e o tempo de equilibração do sistema costuma ser maior que em proteínas simuladas apenas em solvente.

7.12. Aplicações de modelos

A aplicabilidade de um modelo 3D está diretamente relacionada com a acurácia com que este foi gerado. Esta acurácia pode ser avaliada pelo grau de similaridade entre as estruturas 3D da proteína predita e da proteína molde, através do cálculo do desvio médio quadrático (RMSD), que mede as distâncias interatômicas. De acordo com sua acurácia, os modelos 3D gerados por métodos teóricos podem ser aplicados em:

i) Estudos de predição funcional e busca por novos alvos moleculares em organismos patogênicos;

ii) Planejamento racional de fármacos

baseado na estrutura do receptor biológico;

iii) Estudos de variação conformacional por dinâmica molecular;

iv) Planejamento de experimentos de mutagênese sítio-dirigida, fornecendo informações sobre possíveis mutações para testar hipóteses funcionais;

v) Simulações de interações entre proteínas;

vi) Auxiliar no refinamento de estruturas resolvidas por cristalografia de raios-X e por experimentos de RMN.

7.13. Conceitos-chave

Bibliotecas de fragmentos: As bibliotecas de fragmentos são construídas a partir de estruturas tridimensionais determinadas experimentalmente, e são específicas para cada sequência alvo. Possuem tamanhos variados uma vez que os fragmentos devem apresentar alta similaridade local com a sequência alvo.

Campos de força: Referem-se à forma e aos parâmetros (ajustáveis) de funções matemáticas usadas para descrever a energia potencial de um sistema de partículas (moléculas e átomos). As funções e seus parâmetros são derivados de estudos experimentais e de cálculos advindos da mecânica quântica, e que tentam descrever fenômenos atômicos como conformação (*e.g.* diedros) e interações de curto e longo alcance de diferentes classes de moléculas.

Decoy: São modelos gerados pelos diversos métodos de predição de estrutura tridimensional de proteínas. Uma vez que os métodos empregados são não determinísticos, cada execução pode resultar em um modelo diferente. Dentre os *decoys*, encontra-se o modelo que melhor representa o que se supõe ser a estrutura nativa da sequência alvo, porém, para sua identificação faz-se necessário realizar uma filtragem.

Estrutura nativa: É a estrutura tridimensional adotada por uma proteína em seu ambiente fisiológico de ação. É a conformação que desempenha o papel biológico da proteína.

Função de energia: Função pela qual se avalia o estado conformacional de uma proteína. A avaliação é feita baseada no valor de energia total do sistema em estudo, que pode ser composta de termos de energia potencial e cinética. O funcional é específico para cada programa e seus termos são baseados em "Campos de Força".

Metaheuristica: É um processo iterativo que otimiza uma solução candidata segundo um critério de avaliação, geralmente baseada na minimização da "Função de Energia". É comum o uso de métodos de otimização não determinísticos, como por exemplo, algoritmos genéticos e *simulated annealing*.

Modelagem comparativa: É uma classe de métodos de predição de estrutura tridimensional de proteínas. A estrutura da sequência alvo é construída a partir de outras estruturas resolvidas experimentalmente (estruturas molde) e que possuem mais de 25% de identidade (ou ditas homólogas) em relação à sequência de aminoácidos da proteína alvo.

Molde ou *template*: É a estrutura tridimensional de uma proteína determinada experimentalmente e que é usada como base para fornecer informações estruturais aos algoritmos de predição de estrutura de proteínas. Seu uso é dependente do nível de identidade/similaridade entre sua sequência de aminoácidos e a da sequência alvo (sequência que se deseja modelar).

Predição *ab initio*: É uma classe de métodos usada para prever a estrutura tridimensional de uma proteína alvo sem o uso de informações estruturais de quaisquer outras proteínas resolvidas experimental-

mente.

Predição de estruturas de proteínas: É a arte de prever para uma sequência de aminoácidos, através de métodos computacionais, sua estrutura tridimensional mais próxima do que se supõe ser sua estrutura nativa.

Predição *de novo*: É uma classe de métodos usada para prever a estrutura tridimensional de uma proteína alvo, a partir de informações estruturais de proteínas resolvidas experimentalmente (estruturas molde) e sem qualquer parentesco com a proteína alvo. Usam, por exemplo, bibliotecas de fragmentos.

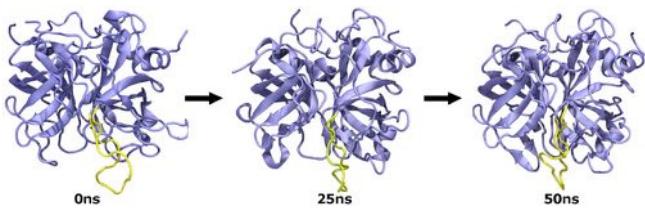
Rotâmeros: São as conformações preferenciais da cadeia lateral de um resíduo de aminoácido. Podem ser combinados em bibliotecas para cada tipo de aminoácido.

Threading: É uma classe de métodos usada na predição de estrutura tridimensional de proteínas e que busca descobrir qual é o tipo de enovelamento mais provável que uma sequência alvo deverá adotar. Esse processo é baseado em estruturas resolvidas experimentalmente (estruturas molde) que não são necessariamente homólogos à proteína alvo.

7.14. Leitura recomendada

CUSTÓDIO, Fábio Lima. ***Algoritmos Genéticos para Predição Ab Initio de Estrutura de Proteínas***. Tese de Doutorado, Laboratório Nacional de Computação Científica: Rio de Janeiro, 2008.

CAPRILES, Priscila Vanessa da Silva Zabala. ***Desenvolvimento e Implementação de um Modelo Coarse-Grained para Predição de Estruturas de Proteínas***. Tese de Doutorado, Laboratório Nacional de Computação Científica: Rio de Janeiro, 2011.

TREVIZANI, Raphael. ***Bibliotecas de fragmentos para a predição de estruturas de proteínas***. Tese de Mestrado, Laboratório Nacional de Computação Científica: Rio de Janeiro, 2008.

LEACH, Andrew R. ***Molecular Modelling Principles and Applications***. 2.ed. Essex: Pearson Education Limited, 2001.

WEBSTER, Davird M. ***Protein Structure Prediction: Methods and Protocols***. Totowa: Humana Press Inc., 2000.

RANGWALA, Huzefa; KARYPIS, George. ***Introduction to Protein Structure Prediction: Methods and Algorithms***. Hoboken: John Wiley & Sons, 2011

8. Dinâmica Molecular

Hugo Verli

Flexibilidade da enzima trombina evidenciada através de simulação por dinâmica molecular.

8.1. Introdução

Figura 1-8: Variação do conteúdo de estrutura secundária da melitina, peptídeo da abelha *Apis mellifera*, como função do tempo. A forma inicial é encontrada no ambiente cristalino, enquanto a final é observada em condições próximas às plasmáticas.

Segundo a IUPAC (*International Union of Pure and Applied Chemistry*), a "dinâmica molecular é um procedimento de simulação que consiste na computação do movimento dos átomos em uma molécula ou de átomos individuais ou moléculas em sólidos, líquidos e gases, de acordo com as leis de movimento de Newton". Em outras palavras, a dinâmica molecular (DM) descreve a variação do comportamento molecular como função do tempo (Figura 1-8).

Quando mencionamos "comportamento molecular", nos referimos a quaisquer propriedades de uma molécula em estudo, tais como seu conteúdo de estrutura $2^{ária}$, orientação de cadeias laterais, conformação de alças e a energia de interação entre diferentes moléculas (enzima e substrato, proteína e proteína, proteína e DNA ou fármaco e receptor). Por outro lado, a ideia de que estas propriedades variam como função do tempo indica que as mesmas não são estáticas, mas se modificam em soluções biológicas. Isto aproxima em muito a DM de métodos experimentais como a Ressonância Magnética Nuclear (RMN, Capítulo 12), que geram medidas representando, de fato, médias temporais, colhidas durante a realização do experimento. Assim, ao final de uma simulação de DM, buscamos estas propriedades médias, representativas de comportamentos biológicos medidos experimentalmente.

A descrição conformacional oferecida pela DM, para uma determinada molécula ou

conjunto de moléculas, baseia-se na solução da 2ª Lei de Newton, onde Fx_i é a força aplicada ao átomo i na posição x, t é o tempo, v a velocidade e a_i a aceleração do átomo i. Por ser baseada na física desenvolvida por Sir. Isaac Newton, a DM faz parte dos métodos denominados Clássicos (também chamados de métodos de mecânica molecular), em oposição aos métodos baseados na física quântica (que deram origem aos denominados métodos de mecânica quântica).

Sir Isaac Newton

$$F_{x_i} = \frac{d^2 x_i}{dt^2} m_i = \frac{\Delta v_i}{\Delta t} m_i = a_i m_i$$

Assim, a DM nos possibilita obter modelos de moléculas muito mais próximos da realidade biológica, pois inclui diretamente características como a flexibilidade molecular (através da variação temporal de propriedades) e a temperatura (através da aceleração dos átomos). A maioria dos fenômenos biológicos estão associados à flexibilidade de biomoléculas, como a catálise e a modulação de canais iônicos e de receptores acoplados à proteína G. De fato, muitos destes processos vêm sendo descritos com sucesso por simulações de DM ao longo dos anos.

Outros tipos de simulação estão disponíveis, tais como o Método de Monte Carlo, a Dinâmica Estocástica e a Dinâmica Browniana. Iremos, contudo, nos ater à DM em decorrência de seu maior uso, nos últimos anos, no estudo de biomoléculas.

Muitos programas (Tabela 1-8) estão disponíveis para a realização de simulações por DM diferindo, por exemplo, quanto a seu acesso (gratuito ou pago), custo computacional (isto é, tempo necessário para a execução de um mesmo cálculo) e tipos de campos de força disponíveis (ver adiante).

8.2. Campos de força

Como visto no item anterior, para descrever a variação da posição x de um átomo i como função do tempo precisamos conhecer o valor da massa de cada átomo, m_i (essa é fácil, vem da tabela periódica) e a força (Fx_i) sobre cada átomo i em uma determinada posição x. A temperatura fornece energia para que os átomos sofram uma aceleração, mudando suas posições no espaço. Contudo,

Tabela 1-8: Alguns dos principais programas disponíveis para simulações por DM.

Programa	Distribuição
Abalone	Gratuito
ADUN	Gratuito
AMBER	Pago
Ascalaph Designer	Gratuito
CHARMM	Pago
Discovery Studio	Pago
GROMACS	Gratuito
GROMOS	Pago
GULP	Gratuito
LAMMPS	Gratuito
MDynaMix	Gratuito
MOE	Pago
MOIL	Gratuito
MOLDY	Gratuito
NAMD	Gratuito
RedMD	Gratuito
TeraQuem	Pago
TINKER	Gratuito
YASARA	Pago

como os átomos não estão isolados, mas ligados a outros átomos formando moléculas que, por sua vez, interagem com outras moléculas, eles estão sujeitos a forças inter-atômicas e inter-moleculares. O cálculo destas forças é realizado por uma outra função matemática, denominada campo de força.

O campo de força, seguindo a definição da IUPAC, pode ser descrito brevemente como "um conjunto de funções e parametrização usadas em cálculos de mecânica molecular". Cada campo de força estabelece um conjunto de equações matemáticas dedicadas a reproduzir aspectos do comportamento molecular, como o estiramento de ligações químicas, a deformação de um ângulo de ligação ou a torção de um diedro, como podemos observar em um espectro de infravermelho. Estas equações, por sua vez, são calibradas (ou seja, parametrizadas) para reproduzir o comportamento dos compostos de interesse (Figura 2-8).

Equações e parametrizações diferentes podem ser empregadas, dando origem a campos de força diferentes, com vantagens e também limitações. Por exemplo, enquanto um tipo de campo de força pode descrever com elevada fidelidade proteínas, ele pode ser bastante limitado na reprodução da geometria de carboidratos ou ácidos nucleicos. Desta forma, ao iniciarmos um estudo por DM, devemos ter em mente qual o tipo de molécula com o qual pretendemos trabalhar e qual o melhor campo de força para descrevê-la.

A escolha de um campo de força não é, contudo, baseada somente no tipo de molécula com o qual queremos lidar. Diversos outros aspectos podem influenciar esta escolha. Existem, por exemplo, diferentes níveis de simplificação na descrição dos átomos (Figura 3-8). O campo de força pode descrever todos os átomos do sistema (em inglês são denominados campos de força *all atom*), mas isto implica em um maior custo computacional, o que pode se tornar proibitivo no estudo de grandes sistemas moleculares se não temos acesso a grandes estruturas de processamento em paralelo (os chamados *clusters*).

Como o elemento encontrado em maior quantidade é o átomo de hidrogênio, uma primeira simplificação é denominada de átomo unido (em inglês são denominados campos de força *united atom*). Neste

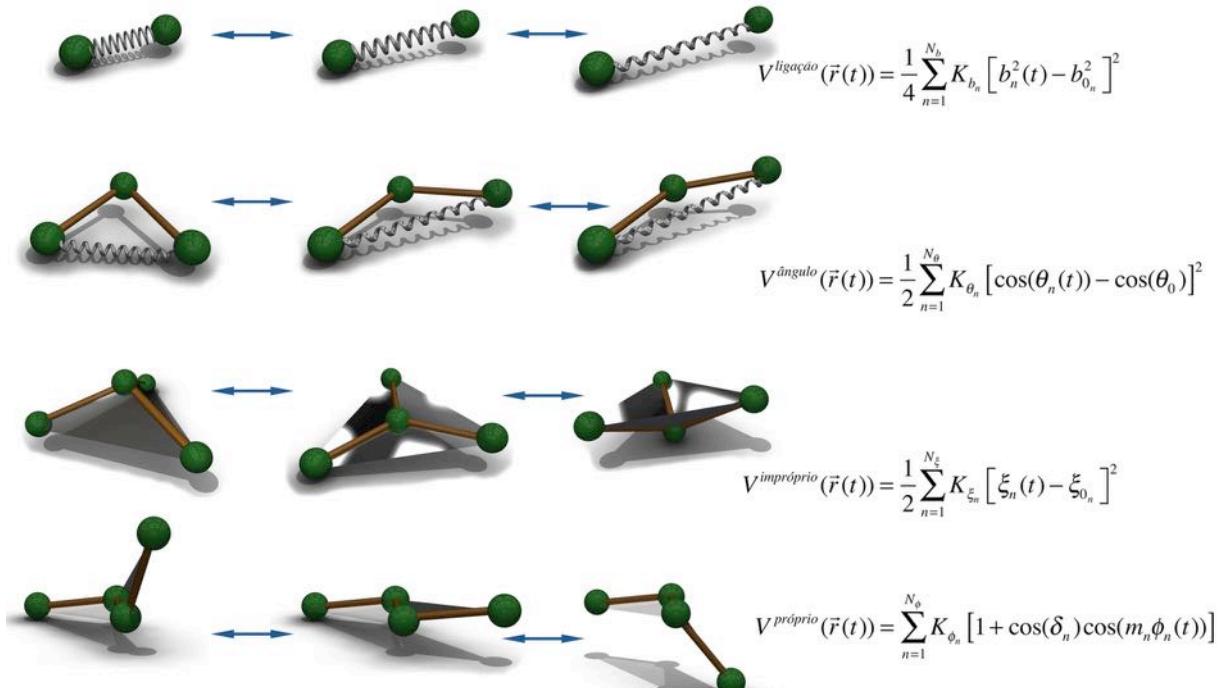



Figura 2-8: Representação de alguns termos que compõem o campo de força GROMOS96. Termos semelhantes são também encontrados em diversos outros campos de força.

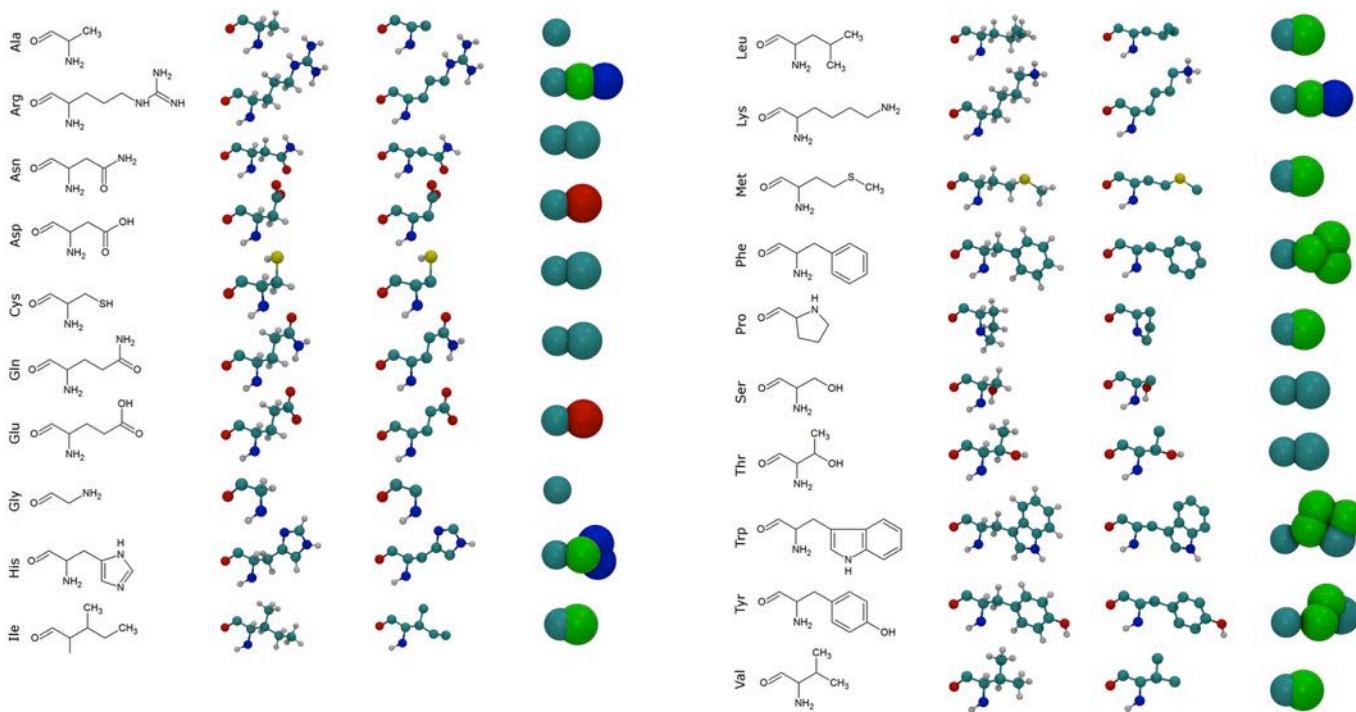



Figura 3-8: Representação dos 20 aminoácidos, codificados no genoma para síntese proteica, em um campo de força descrevendo todos os átomos, em um campo de força de átomo unido e *coarse-grained*.

caso, os átomos de hidrogênio apolares, ou seja, aqueles ligados a átomos de carbono, são unidos a este elemento, dando origem a um pseudoátomo representando as propriedades de grupos CH, CH_2 ou CH_3. Exceção se dá para o grupo CH de anéis aromáticos, que tem os átomos de hidrogênio descritos explicitamente nos campos de força de átomo unido mais modernos, como o GROMOS96.

Há, por fim, um terceiro nível de simplificação, denominado *coarse-grained* (CG). Neste campo de força, vários átomos podem ser agregados em uma única partícula, análoga ao pseudoátomo do modelo de átomo unido. Por exemplo, todo um aminoácido pode ser considerado como uma única partícula, como é o caso da alanina e da glicina no campo de força MARTINI. Em outros resíduos, este campo de força considera o esqueleto peptídico como uma partícula e a cadeia lateral de uma (como na cisteína, treonina e serina) a três (histidina e fenilalanina) ou quatro (triptofano) partículas.

Quanto maior a simplificação, menor custo computacional do cálculo. Em outras palavras, podemos simular sistemas com maior número de átomos por mais tempo em computadores mais baratos. Infelizmente, estas simplificações trazem consigo algumas limitações. No caso do CG, perde-se a capacidade de descrever elementos de estrutura $2^{ária}$, mantendo-se somente a forma global da molécula em estudo. Assim, em estudos onde são esperadas mudanças no conteúdo de estrutura $2^{ária}$ o método de CG não é indicado. Mas, por ser muito rápido, pode descrever movimentos entre diferentes domínios de uma dada proteína, o que é difícil de ser observado, usualmente, nos demais campos de força. Por outro lado, o caso dos modelos de átomo unido traz limitações como a dificuldade em se utilizar estes campos de força na obtenção e refinamento de modelos 3D de macromoléculas a partir de dados de RMN (Capítulo 12).

Outra diferença entre os campos de força diz respeito à descrição das moléculas de água, o principal solvente de biomoléculas (Tabela 2-8). De fato, uma das grandes vantagens do método de DM é a capacidade de incluir a presença de moléculas de água nos modelos gerados, descrevendo as suas interações, como função do tempo, com os compostos em estudo. Da mesma forma que visto para os campos de força, existem diversos modelos para descrição de moléculas de água, por vezes com mais de uma opção para um mesmo campo de força.

Estes organizam-se em dois grandes grupos: os modelos explícitos e os implícitos.

Tabela 2-8: Alguns dos modelos de água mais comumente empregados em simulações por DM[a].

Modelo	Campos de força onde são empregados	Tipo
SPC SPC/E TIP3P	AMBER, GROMOS, OPLS	
TIP4P TIP5P	AMBER, CHARMM, OPLS	
MARTINI	Martini	

[a]Uma revisão mais completa pode ser encontrada no site: www1.lsbu.ac.uk/water/models.html

Enquanto os modelos explícitos incluem os átomos da molécula de água, fisicamente, na simulação, os modelos implícitos (também chamados de modelos contínuos ou *continuum models*) não incluem estas moléculas diretamente, mas indiretamente, através da representação das propriedades dielétricas do solvente. Os átomos que compõem a água não participam das simulações, tornando o cálculo extremamente rápido (usualmente, a grande maioria dos átomos em um sistema a ser simulado por DM se refere ao solvente). Infelizmente, enquanto estes modelos implícitos são bastante eficientes no estudo de proteínas e ácidos nucleicos, o mesmo não vem se mostrando para carboidratos, compostos altamente polares que interagem intensamente com o solvente.

Embora os principais campos de força empregados atualmente (AMBER, CHARMM, OPLS e GROMOS) sejam compostos por equações bastante semelhantes (ver a seguir), cada um foi construído a partir de decisões metodológicas distintas apresentando, portanto, particularidades importantes. Como consequência, normalmente os parâmetros de um campo de força não são transferíveis para outro campo de força.

A importância de conhecermos estas características, reconhecendo cada campo de força como entidade única, reside no fato de que um grande número de compostos de interesse biológico não é descrito nos parâmetros atuais, o que pode limitar o seu estudo computacional. Dentre estes compostos com carências de parâmetros podemos citar aminoácidos modificados (além dos 20 codificados no genoma), neurotransmissores, hormônios, fosfolipídeos, carboidratos, produtos naturais e, por fim, fármacos. Como simulações por DM podem ser cálculos extremamente demorados, deixar para descobrir no meio do trabalho que seu modulador de interesse não tem parâmetros no campo de força escolhido pode lhe custar alguns meses de trabalho.

Em linhas gerais, tanto a distância entre 2 átomos ligados quanto o ângulo entre 3 átomos consecutivos é descrita a partir de $V_{ligação/ângulo} = K_n [n - n_o]^2$, onde V é a energia, n é a distância ou ângulo em um dado momento, n_o é a distância ou ângulo de referência e K_n é a constante de força da mola que mantém esses valores ao redor dos valores de referência (Figura 2-8).

Para diedros, a função mais usualmente empregada é baseada em $V_{diedro} = K_\chi [1 + cos(n_\chi - \delta)]$, sendo V a energia, $_\chi$ o valor do diedro e K_χ a altura da barreira de energia entre diferentes estados conformacionais. Estes estados surgem porque um diedro pode rodar 360º e, ao longo desta rotação, apresentar múltiplos mínimos de energia. Assim não há, necessariamente, uma única geometria de referência. O perfil rotacional dos diedros tem a adição do parâmetro n, que descreve a multiplicidade do diedro (ou seja, o número de mínimos de energia) e δ, que diz respeito à mudança de fase e à localização do máximo de energia ao longo do perfil da rotação do diedro.

Apesar da semelhança nesses termos, existem diferenças importantes que devem ser consideradas. O CHARMM, por exemplo, emprega uma equação adicional na descrição dos ângulos de ligação, chamada

Urey-Bradley, que busca preservar a distância entre o primeiro e o terceiro átomos de um ângulo. Outra diferença se refere aos termos que descrevem a planaridade ou quiralidade em um conjunto de quatro átomos, o que é usualmente chamado de diedro impróprio (Figura 2-8). Enquanto AMBER e OPLS os descrevem da mesma forma que os demais diedros (também chamados de diedros próprios), CHARMM e GROMOS aplicam uma equação diferente, que se assemelha àquela empregada para distâncias e ângulos.

Abordar com profundidade a construção de parâmetros para campos de força está além do objetivo deste livro. Mas em muitos casos há uma solução um pouco mais simples para o problema. Uma característica importante de campos de força é a chamada transferabilidade. Isto significa que grupos químicos semelhantes possuem propriedades semelhantes que podem, assim, serem transferidas de uma molécula para outra. Por exemplo, o grupo hidroxila de um resíduo de Ser é equivalente ao grupo hidroxila de um resíduo de Thr. Assim, há uma redução enorme na necessidade de construção de parâmetros para novos compostos, se respeitarmos a semelhança química entre eles.

8.3. Minimização de energia

Quando iniciamos um estudo baseado em simulações por DM, podemos empregar estruturas de partida de diferentes origens, como modelos teóricos (ver capítulo 7) ou ainda dados experimentais de cristalografia de raios-X (ver capítulo 13) ou de RMN (ver capítulo 12). Independente de sua origem estas estruturas, ao serem solvatadas, criam interações soluto-solvente até então inexistentes (seja pelo dado ser teórico obtido no vácuo, em ambiente cristalino ou como uma média de diferentes conformações). Mas o solvente precisa se adaptar ao redor de seu soluto, e isto precisa ser corrigido antes que a simulação por DM se inicie. Por exemplo, quando o programa insere uma molécula de água, esta pode ter seu hidrogênio apontando para um átomo de hidrogênio da cadeia lateral de uma arginina, promovendo uma repulsão eletrostática pela proximidade de duas cargas de sinais iguais. Se isto não for corrigido antes do início da DM, a liberação desta energia na simulação pode gerar uma explosão da simulação (Figura 4-8) ou, de forma mais sutil (mas nem por isso menos perigosa para o estudo), promover mudanças conformacionais na proteína, ou mesmo desnaturações. Em outros casos, como na obtenção de modelos teóricos para a estrutura 3D de proteínas, a construção de cadeias laterais de aminoácidos pode aproximá-las artificialmente (e excessivamente) de outros resíduos.

Assim, uma das principais formas de tentar eliminar estes problemas reside no cálculo de minimização de energia (Figura 5-8). Durante este cálculo, a energia global do sistema é reduzida, alcançando por fim uma conformação mais estável para o sistema em estudo (ou seja, um estado de mínimo de energia).

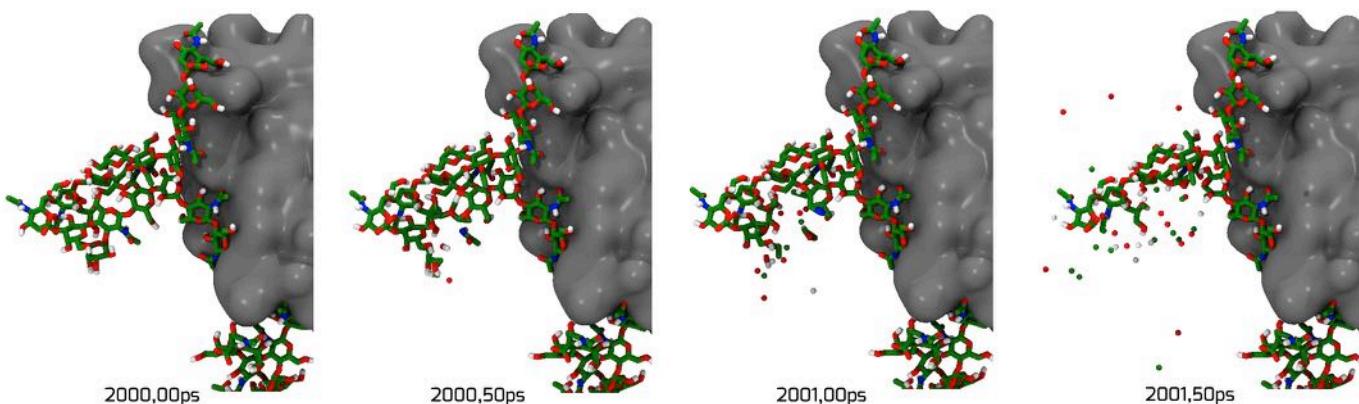



Figura 4-8: Explosão em uma simulação por DM.

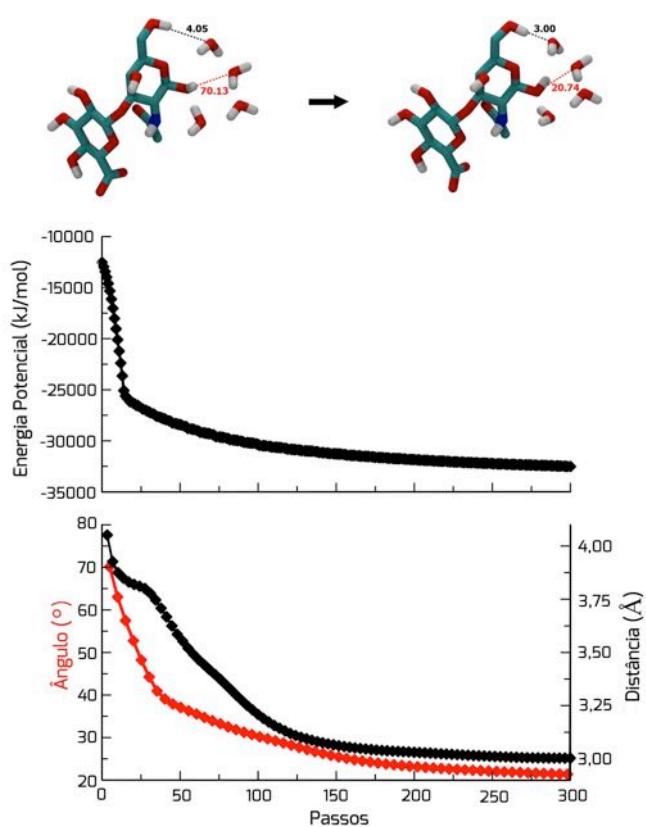



Figura 5-8: Exemplo da evolução de propriedades moleculares no decorrer de uma minimização de energia. A cada passo, a energia do sistema diminui, com a redução de contatos desfavoráveis e a formação de interações intra- e inter-moleculares como ligações de hidrogênio.

8.4. Simulações por DM

Além da escolha do campo de força e do modelo de água, o preparo e a análise de uma simulação por DM deve considerar alguns aspectos metodológicos importantes, dentre os quais destacaremos as condições periódicas de contorno, a equilibração, a amostragem, o tempo de integração e o cálculo de interações não ligadas. Uma escolha inadequada destas propriedades pode significar desde um maior custo computacional (isto é, uma simulação demorando mais do que precisaria) a resultados que não representam situações reais.

Condições periódicas de contorno

Quanto maior o número de moléculas incluídas em uma simulação, maior será o tempo necessário para realizar o cálculo. Por isso, buscamos sempre incluir o menor número de moléculas possível capaz de descrever as condições experimentais ou fisiológicas de referência. No caso da proteína, estamos na maioria das vezes ainda limitados a simulação de uma única molécula (salvo no caso de oligômeros). Contudo, a proteína não costuma ser a parte mais cara computacionalmente do cálculo, mas sim a inclusão do solvente (explícito). Uma otimização no número de moléculas de água pode representar uma grande otimização no tempo de máquina para conclusão da simulação (o que permite aumentar o tamanho da amostragem do estudo, ver adiante).

Uma forma de controlar o número de moléculas de água é controlando o tipo de "caixa" onde o sistema será simulado. Por caixa entendemos o espaço tridimensional onde soluto (biomolécula) e solvente (normalmente água) são colocados. O tamanho e a forma desta caixa, usualmente centralizada no soluto, definirá a quantidade de solvente a ser inserida.

Atualmente, não é comum definir a forma da caixa como uma esfera, por motivos que explicaremos a seguir. As formas mais comuns são cúbica, octaédrica e dodecaédrica. A forma de um octaedro apresenta 77% do volume de um cubo, enquanto que o dodecaedro 71%, representando a forma mais próxima de uma esfera. Contudo, como a forma de proteínas e outras biomoléculas varia muito, devemos avaliar qual caixa se adequa melhor ao sistema em estudo. Por exemplo, a simulação de membranas é normalmente realizada em um cubo ou uma forma retangular, que pode ser uma boa alternativa também para proteínas em forma de bastão.

O uso de uma caixa em forma de esfera ao redor da proteína de interesse nos levaria a um aproveitamento do espaço tridimensional melhor do que o dodecaedro, economizando mais moléculas de água e, assim, liberando custo computacional. Contudo, as moléculas em uma simulação por DM podem se difundir ao longo da caixa. Como além da caixa de simulação temos condições de vácuo, o solvente iria progressivamente evaporar, a partir da face da esfera. A forma de

impedir isso é criar uma força que impeça as moléculas do sistema de ultrapassarem os limites desta esfera, o que representa a inclusão de forças artificiais, não observáveis em condições biológicas.

As formas geométricas empregadas mais frequentemente em em simulações por DM estão relacionadas a uma estratégia denominada condições periódicas de contorno (Figura 6-8). Estas formas permitem que uma caixa de simulação seja replicada em todas as suas dimensões, de forma periódica. Estas réplicas são idênticas à caixa construída, de forma que um movimento molecular em uma será idêntico ao movimento da mesma molécula na outra. Mas, agora, a face da caixa não está em contato com o vácuo, mas com solvente. E, caso uma molécula saia da caixa central, uma de suas imagens entrará pela face oposta, mantendo o número de moléculas constante. Isto representa uma continuidade da solução, nos aproximando de condições experimentais.

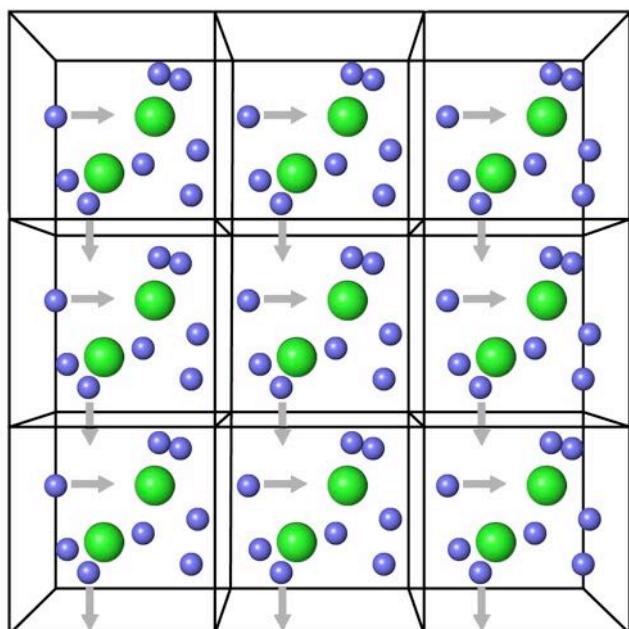



Figura 6-8: Representação das condições periódicas de contorno em uma simulação por DM. Somente a caixa central é simulada, enquanto que as réplicas garantem a continuidade do sistema, isto é, ausência de contato das moléculas com o vácuo.

Devemos, contudo, tomar cuidado para não definir uma caixa excessivamente pequena, buscando economizar custo computacional ao reduzir a quantidade de solvente excessivamente. Se a caixa for pequena demais, a proteína pode interagir com suas imagens, geradas pelas condições periódicas de contorno, criando uma situação artificial que provavelmente irá deturpar os resultados obtidos. É importante, assim, avaliar se o corte das interações não ligadas (ver adiante) é menor que a distância da proteína às suas imagens.

Equilibração

A ideia de equilibração de uma simulação por DM se refere à estabilização de suas propriedades, ou seja, que estas alcancem um estado de equilíbrio. Considera-se que, antes de estarem equilibradas, as propriedades em estudo apresentam variações ou comportamentos não representativos das situações de interesse. Assim, é necessário que o tempo de simulação seja suficientemente longo (tamanho da amostragem, ver adiante) para que as propriedades em estudo estejam adequadamente equilibradas. Na Figura 1-8, por exemplo, a simulação de um monômero de melitina demora em torno de 4 ns para se equilibrar.

Um dos motivos mais comuns para a necessidade de equilibração é devido ao uso de estruturas 3D derivadas de ambientes cristalinos, isto é, aquelas obtidas por cristalografia de raios-X. Este ambiente apresenta concentração de proteínas muito maior do que aquela observada, usualmente, nas condições biológicas de interesse, por vezes em estados oligoméricos não observados em condições biológicas. Assim, a remoção destes contatos e sua substituição por moléculas de água, acarretará em uma instabilidade inicial na simulação, envolvendo: 1) a perda de contatos cristalográficos, e 2) a formação de interações com moléculas de água.

Infelizmente, a busca por tempos de simulação "suficientemente longos" para equilibração das propriedades de interesse pode ser desafiadora, pois nem todas as propriedades moleculares equilibram a uma mesma velocidade. Por exemplo, a interação de uma proteína com o solvente equilibra usualmente mais rapidamente do que a perda ou a formação de estrutura 2^{ária}. Estas, por sua vez, equilibram mais

rapidamente que o movimento de domínios em uma dada proteína.

Amostragem

A amostragem de uma simulação por DM se refere a quão bem ela é capaz de descrever o comportamento do sistema molecular em estudo. Idealmente, a amostragem de uma simulação deve ser longa o bastante para descrever os fenômenos de interesse. Contudo, a simulação de sistemas complexos como aqueles envolvendo biomoléculas frequentemente esbarra em amostragens ainda inalcançáveis em decorrência de seu elevado custo computacional.

A maneira mais simples de se entender a amostragem é considerando o tamanho da simulação em uma escala de tempo. Um maior tempo de simulação implica em uma maior amostragem. Contudo, diversos aspectos podem interferir neste entendimento. O aumento do número de moléculas e átomos no sistema aumenta o número de possíveis conformações a serem adotadas. Por outro lado, o uso de campos de força do tipo átomo unido ou ainda *coarse-grained*, ao reduzir o número de átomos, reduz o número de possíveis estados conformacionais a serem adotados pelo sistema, tornando assim a amostragem maior em uma mesma escala de tempo.

Tempo de integração

O cálculo de uma simulação por DM não gera informações contínuas, mas sim é dividida em pequenos passos, usualmente na escala de femtossegundos (fs). A sucessão destes passos dará origem ao nosso entendimento de trajetória, isto é, à evolução temporal do comportamento molecular na simulação realizada. O tamanho destas partes é o que chamamos de tempo de integração (Figura 7-8).

A definição de um valor apropriado para o tempo de integração está diretamente relacionada ao tamanho da amostragem da simulação e, por conseguinte, ao custo computacional da mesma. Conforme ilustrado na Figura 7-8, a descrição de uma determinada propriedade tempo-tempendente

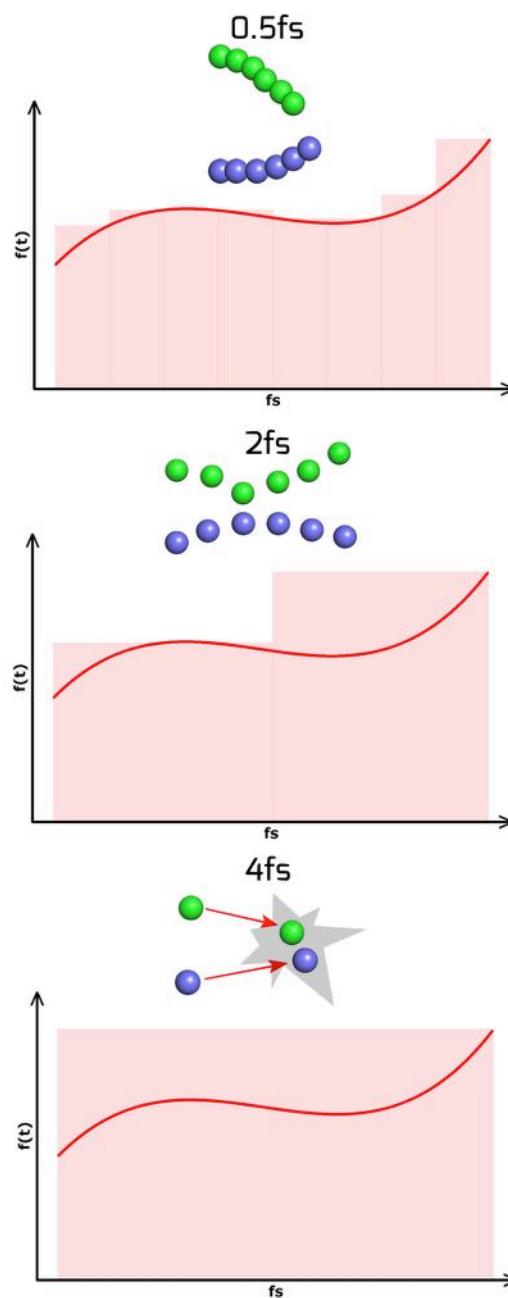



Figura 7-8: Representação do efeito de diferentes tempos de integração na amostragem de uma simulação por DM. Valores muito pequenos (0,5fs) descrevem fenômenos com maiores detalhes, mas mais lentamente. Valores muito grandes (4,0fs) apresentam menores custos computacionais, mas podem dar origem a instabilidades.

pode ser feita empregando-se diferentes valores de tempo de integração. Quanto maior este valor, menos passos de cálculo serão necessários à descrição do fenômeno e, por conseguinte, menor será o custo computacional associado. Quanto menor este valor,

mais passos serão necessários e, assim, maior o custo computacional. Infelizmente, o uso de tempos de integração muito elevados pode gerar instabilidades na trajetória, de forma que valores intermediários são usualmente empregados, no caso da Figura 7-8, 2fs.

Os valores de tempo de integração mais frequentemente empregados em simulações baseadas em campos de força atomísticos (isto é, todos os átomos são descritos) ou de átomo unido são 1fs, 2fs ou 5fs. O uso de 1fs é realizado quando as moléculas e suas ligações são tratadas como flexíveis durante a simulação, enquanto 2fs requerem o tratamento das ligações químicas como rígidas. Já para o uso de 5fs, toda a molécula é tratada como rígida (ou seja, ângulos e diedros não podem ser modificados), uma alternativa pouco utilizada no estudo de sistemas biológicos. Em algumas situações podem ser empregados tempos de integração menores que 1fs, mantida toda a flexiblidade da molécula. Em outros casos, como em simulações do tipo *coarse-grained*, tempos de integração de até 40fs.

Cálculo de interações não ligadas

Uma das partes mais custosas computacionalmente em simulações por DM envolve o cálculo das interações não ligadas, isto é, interações eletrostáticas (calculadas por termos de Coulomb) e de van der Waals (calculadas pelo potencial de Lennard-Jones). Para se ter uma ideia, enquanto o número de termos ligados (isto é, ligações, ângulos e diedros) é proporcional ao número de átomos, o número de interações não ligadas aumenta como função do quadrado do número de átomos do sistema. Assim, economizar custo computacional no cálculo destas interações representa uma significativa redução no custo da simulação como um todo. Como estas interações decrescem rapidamente em intensidade conforme dois átomos se distanciam no espaço, é possível realizar cortes nestas interações (*cut-off*). Em outras palavras, a partir da distância definida por estes cortes, nenhuma interação não ligada será calculada (Figura 8-8).

Por exemplo, consideremos dois possíveis raios de corte na simulação do soluto apresentado na Figura 8-8. O uso do raio **a** representaria um menor custo computacional, tendo em vista que nenhuma interação de Coulomb seria avaliada a partir desta distância. Já o uso do corte **b** traria um maior custo computacional, incluindo as interações entre o soluto e as moléculas na faixa cinza da figura. Contudo, ao reduzir o custo computacional, o corte **a** potencialmente implicará na perda de informações importantes, por ser muito próximo do soluto. Assim, a distância **b** seria preferível.

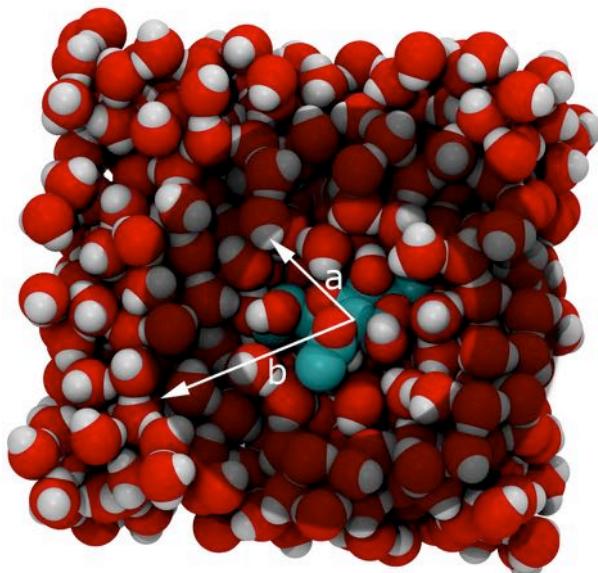



Figura 8-8: Representação de regiões de corte, a e b, a partir de um soluto, para cálculo de interações não ligadas.

A eliminação repentina da avaliação das interações não ligadas através de um *cut-off* pode gerar instabilidades ou erros na amostragem da simulação. Desta forma, estas interações a longas distâncias costumam ser descritas por outros tipos de métodos, como PME, Ewald ou Campo de Reação (*Reaction-Field*), dentre outros. Este tratamento é usualmente aplicado somente às interações de Coulomb, mais sensíveis a efeitos originados de cortes nas interações.

8.5. Estratégias de análise

Um dos maiores desafios em um estudo baseado em DM frequentemente reside mais na análise e interpretação dos resultados obtidos do que no preparo do sistema. De fato, simulações de proteínas em água podem gerar facilmente muitas dezenas de gigabytes de dados. Como retirar informações destas trajetórias, quais informações retirar e como interpretar estas informações, no contexto do

assunto em estudo, envolvem muitas vezes mais tempo do que a simulação computacional em si.

Os tipos de análises a serem empregadas estarão intrinsecamente relacionados à natureza do problema em estudo. Por exemplo, se estamos estudando uma proteína tentando mimetizar o ambiente nativo da mesma, em princípio, ela não pode se desnaturar durante a simulação. Por outro lado, o estudo de membranas elimina esta preocupação mas nos traz a necessidade de avaliar as propriedades dos lipídeos enquanto imersos num fluido. Adicionalmente, dados prévios sobre características estruturais e/ou funcionais das moléculas em estudo, obtidos tanto por métodos computacionais quanto por outras ferramentas experimentais são fundamentais na concepção, preparo, execução e análise de estudos por DM. Esta é, fundamentalmente, a razão pela qual este livro traz em si diversos métodos experimentais.

Neste momento, a adequação da amostragem às propriedades em estudo assume importância fundamental. Se buscamos estudar o movimento de domínios de uma proteína, simulações de dezenas de nanossegundos não serão suficientes, requerendo potencialmente tempos próximos de microssegundos, possivelmente inviabilizando o estudo por DM. De forma semelhante, a observação do enovelamento de proteínas por DM é impraticável na grande maioria dos casos, salvo em pequenas proteínas ou peptídeos, de qualquer forma, requerendo no mínimo centenas de nanossegundos. Por outro lado, reorientação ou refinamento de cadeias laderais de resíduos de aminoácidos ou de ligantes em complexos fármaco-receptor podem ser observados frequentemente em algumas dezenas de nanossegundos.

As análises de simulações por DM devem, preferencialmente, ser realizadas observando propriedades de complexidade crescente (o que costuma estar associado ao tempo requerido à equilibração desta propriedade). Assim, as primeiras propriedades a serem avaliadas são normalmente a pressão (no caso de simulações NPT, mais comuns em sistemas biológicos), o volume (no caso de simulações NVT), a densidade e a energia total do sistema. Todas estas propriedades devem alcançar um patamar estável, paralelo ao eixo x (tempo). Pode-se observar alguma variação no início da simulação mas, em seguida, devem atingir este patamar e se manter neste nível ao longo da simulação. Estas costumam ser propriedades de rápida equilibração em simulações por DM.

Garantidas estas propriedades, podemos passar à análise de aspectos mais complexos, como do comportamento da estrutura proteica ao longo da simulação. Neste grupo, as ferramentas mais comumente empregadas incluem o RMSD, o RMSF, o raio de giro, distâncias entre átomos ou grupamentos e a evolução do conteúdo de estrutura 2^{ária} como função do tempo.

O RMSD (do inglês *root mean square deviation* ou desvio quadrático médio) é uma das principais estratégias de análise empregadas no estudo por DM de proteínas (Figura 9-8A). Indica o quanto a estrutura da proteína de interesse se modifica ao longo de uma simulação, em relação à estrutura de partida, normalmente cristalográfica. Assim, é usual que haja um aumento progressivo no RMSD de uma proteína, partindo de 0, até um patamar, o que pode indicar a equilibração do sistema. Este patamar pode variar em função das características da proteína mas, como um ponto de partida, podemos considerar um valor em torno de 3 Å quando todos os átomos do sistema são empregados na medida. Valores acima deste podem sugerir movimentos maiores de alças, em relação ao cristal, ou perda de estrutura 2^{ária}, enquanto valores menores tendem a indicar sistemas mais semelhantes à referência cristalográfica.

Uma consideração importante quando realizamos análises de RMSD se refere ao fato de que esta análise oferece uma medida média de um conjunto de átomos, selecionados para a análise. Se todos os átomos de uma proteína são considerados, como no exemplo acima, os valores observados trazem consido influências de diferentes regiões da proteína. Por exemplo, normalmente conjuntos de hélices α se modificam menos durante uma simulação do que regiões de alças. Caso façamos uma análise de RMSD separada para estas regiões, veremos hélices α com valores menores e alças com valores maiores do que aqueles considerando

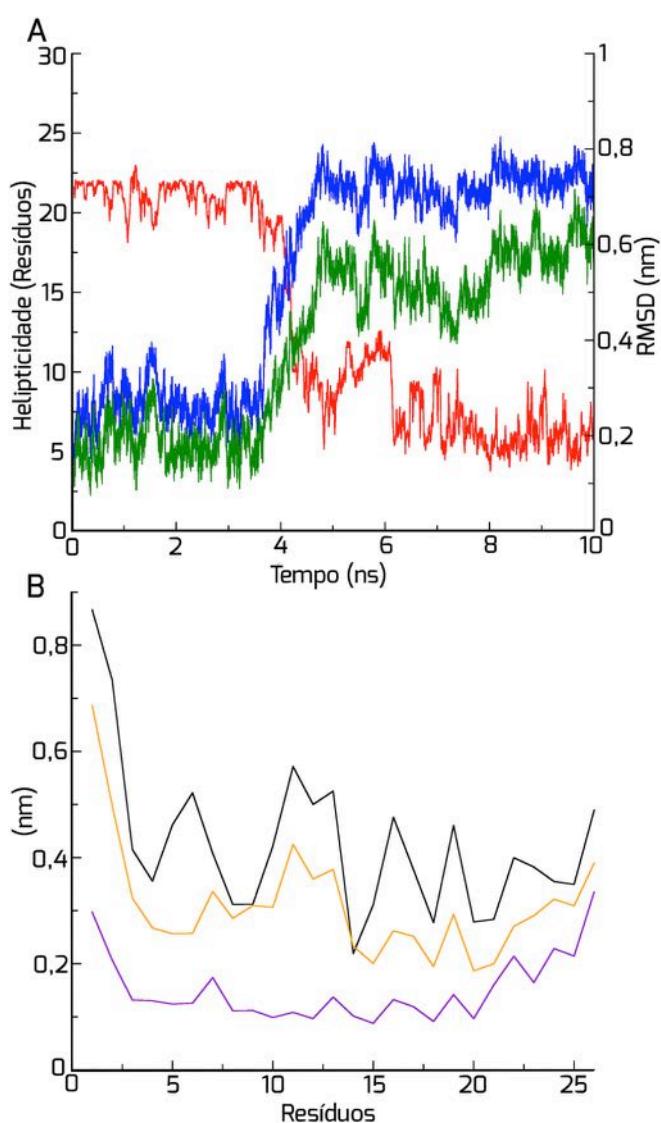



Figura 9-8: A) Helipticidade (vermelho) e RMSD, e B) RMSF para a melitina. O RMSD foi calculado para toda a proteína (azul) e para o esqueleto peptídico (verde). Já o RMSF foi medido como média para toda a trajetória (preto), para os primeiros 3 ns (roxo) e para os últimos 5 ns (laranja).

ambas regiões juntas. Processo similar ocorre caso consideremos todos os átomos do sistema (maior RMSD) ou simplesmente o esqueleto peptídico (menor RMSD) (Figura 9-8A).

Na análise por RMSD, todo resultado obtido irá depender da geometria de partida da simulação, usualmente cristalográfica. O RMSF (do inglês *root mean square fluctuation* ou flutuação quadrática média), em contrapartida, não apresenta esta dependência, mas descreve a variação da posição dos átomos (ou resíduos de aminoácidos) durante a simulação, indicando a flexibilidade do sistema (Figura 9-8B). Valores maiores de RMSF serão, portanto, usualmente observados para alças, e valores menores para hélices α. Por outro lado, regiões de hélices α apresentando valores elevados de RMSF podem estar passando, durante a simulação, por perda de sua estrutura $2^{ária}$.

Enquanto o RMSD apresenta um valor médio, a cada passo da simulação, para todos os átomos do sistema, o RMSF apresenta um valor médio, para cada átomo ou resíduo (usualmente mais útil para proteínas), ao longo de todos os passos da simulação. Assim, valores de RMSF para toda a trajetória podem diferir, por exemplo, daqueles observados no início e/ou no final da simulação (Figura 9-8B).

Ainda, ao observarmos o quanto uma proteína muda sua forma 3D em relação ao cristal ou a flexibilidade de cada resíduo ao longo da simulação, não temos informações diretas sobre o comportamento dos elementos de estrutura $2^{ária}$ da proteína. Um valor de RMSD elevado pode tanto sugerir a desnaturação de uma hélice quanto uma reorientação da mesma que, contudo, pode se manter enovelada. Da mesma maneira, um resíduo muito flexível (conforme observado pelo RMSF) não necessariamente será encontrado somente em alças. Para tal, devemos empregar análises específicas capazes de indicar como a estrutura $2^{ária}$ da proteína se comporta na simulação por DM.

Conforme observado no Capítulo 2, a definição da estutura $2^{ária}$ não é algo tão simples e direto como possa parecer. Existe mais de uma forma de definir hélices e folhas, e diferentes estratégias podem oferecer resultados distintos. Por exemplo, o programa DSSP descreve a estrutura $2^{ária}$ a partir do padrão de ligações de hidrogênio na sequência polipeptídica. À informação relacionada a interações por ligação de hidrogênio o programa STRIDE adiciona parâmetros torsionais relacionados ao esqueleto peptídico.

Outro aspecto importante quanto à análise do comportamento da estrutura $2^{ária}$ diz respeito à escala de tempo na qual hélices e fitas se enovelam. Enquanto hélices usualmente se enovelam numa escala de tempo de centenas de nanossegundos, simulações de poucas dezenas de nanossegundos terão dificuldades em prever estes fenômenos. O caso de fitas é ainda mais complexo, exigindo escalas de tempo uma ordem de grandeza superiores.

Uso de estatística

Embora seja prática corriqueira, mesmo obrigatória, na grande maioria dos métodos experimentais empregados no estudo de sistemas biológios, o uso de métodos estatísticos não é, ainda, comum na análise de resultados obtidos em simulações por DM. Isto se deve ao fato de que, em uma mesma simulação, são normalmente gerados centenas de milhares ou mesmo milhões de dados para uma mesma variável (tamanho da simulação dividido pelo tempo de integração). O grande *n* assim obtido tenderá a tornar estatisticamente significativa mesmo variações bem pequenas nas propriedades de interesse.

Com a redução no custo dos computadores e aumento em sua velocidade, assim como na melhoria dos programas disponíveis, uma nova abordagem vem se apresentando, aproximando a análise de simulações por DM de estudos experimentais convencionais. Trata-se da realização de múltiplas simulações para um mesmo sistema. Assim, a informação a ser empregada nas análises é a média da informação gerada nas diversas simulações.

8.6. Limitações atuais da DM

Como toda técnica experimental, simulações por DM possuem limitações importantes que devem ser conhecidas pelos seus usuários de forma a reduzir a chance de interpretações equivocadas dos resultados obtidos.

Uma consequência direta da realização de cálculos baseados na mecânica molecular, ou seja, empregando campos de força, é a ausência de elétrons. Este tipo de cálculo não considera os elétrons e, por conseguinte, os resultados obtidos apresentam limitações em lidar com fenômenos envolvendo elétrons diretamente. Assim, simulações por DM não são capazes, por exemplo, de descrever reações químicas, como as observadas na ação de enzimas ou em processos de oxidação e redução. Uma alternativa recente para esta limitação envolve métodos denominados híbridos entre a mecânica molecular e a mecânica quântica.

Simulações por DM apresentam grande dificuldade em descrever a energia livre de Gibbs associada a eventos moleculares. Portanto, informações sobre constantes de equilíbrio, constantes catalíticas ou afinidades entre moléculas não são usualmente acessíveis, com precisão, através destas técnicas. Embora diversas técnicas gerem estimativas de energia livre associadas à DM, como a perturbação da energia livre, o *linear interaction energy* e a metadinâmica, cada uma possui suas próprias limitações, dificultando seu uso amplo em estudos por DM.

Por fim, e não menos importante, temos a dificuldade em obter amostragens compatíveis com fenômenos observáveis em experimentos ou fisiologicamente. Mesmo nos maiores centros de supercomputação do mundo, ainda não chegamos, na grande maioria dos casos, em escalas de tempo compatíveis com o comportamento de proteínas em soluções biológicas. Por isso, devemos ter em mente que os resultados obtidos, por mais confiáveis e corretos que sejam, não necessariamente representam, estatisticamente, fenômenos medidos em solução.

8.7. E outras biomoléculas?

A maior parte da literatura, seja em livros seja em artigos, se refere ao estudo de proteínas. Ácidos nucleicos, membranas e carboidratos vêm sendo estudados com menos frequência, comparativamente, ao longo dos anos. Embora possa se justificar esta diferença em decorrência do fato de que as proteínas são as moléculas efetoras da informação genética, esta não é a única justificativa, tampouco proteínas são os únicos compostos biológicos importantes para a manutenção da vida.

O estudo de moléculas de DNA, por exemplo, vem ganhando importância com o desenvolvimento de compostos capazes de interagir, seletivamente, com regiões específicas do DNA, como é o caso dos agentes antineoplásicos. Enquanto moléculas de DNA apresentam estruturas mais ou menos bem definidas, moléculas de RNA são extremamente versáteis e complexas conformacio-

nalmente, a cada momento se mostrando como capazes de atuarem em mais fenômenos biológicos. Valorização semelhante vem sendo observada para membranas e carboidratos que, progressivamente, deixam de ter papéis passivos, simplesmente estruturais, passando a desempenhar papéis ativos, sinalizando diretamente múltiplas respostas em organismos.

Assim, a construção de modelos computacionais para o estudo de biomoléculas deve incluir o máximo de propriedades importantes ao desenvolvimento normal de suas funções, em condições nativas. Uma proteína inserida em membrana irá exigir a inclusão da membrana nas simulações, da mesma maneira que uma glicoproteína irá demandar a inclusão da parte sacarídica em seu estudo.

Do ponto de vista da disponibilidade de parâmetros de campos de força, diferentes classes de biomoléculas apresentam diferentes disponibilidades de parâmetros. Por isso, é importante considerar todos os componentes do sistema molecular quando da escolha do campo de força a ser empregado. Se a nossa molécula em estudo é uma glicoproteína, não adianta empregar um campo de força excelente para carboidratos se o mesmo não possui parâmetros para o estudo de proteínas.

Atualmente, os principais campos de força são capazes de descrever a grande maioria das classes de biomoléculas. Originalmente, no entanto, o campo de força AMBER foi desenvolvido para o estudo de ácidos nucleicos e proteínas, o CHARMM para proteínas, o GROMOS para lipídeos e o OPLS para líquidos e solventes. Com o passar do tempo, cada um desses parâmetros foi sendo aprimorado focando em diferentes biomoléculas, de forma que, hoje, alguns são empregados com maior freqüência para determinados sistemas por melhor descreverem suas propriedades (estruturais, conformacionais ou físico-químicas).

No caso específico de proteínas, os campos de força citados acima descrevem de forma semelhante sua estrutura, conformação e dinâmica. No caso de lipídeos, a maior parte dos estudos envolve os campos de força CHARMM e GROMOS, embora o último ofereça um ganho de velocidade de até nove vezes devido a sua natureza de átomo unido.

Para ácidos nucleicos, os campos de força mais amplamente utilizados são o AMBER e o CHARMM, tanto para DNA quanto para RNA.

A parametrização de carboidratos, por sua vez, está imersa em desafios devido à sua elevada complexidade estrutural e conformacional, de forma que uma sucessão de novos parâmetros vêm sendo desenvolvida.

Por fim, o grupo de compostos mais desafiadores com relação à disponibilidade prévia de parâmetros envolve os fármacos ou moduladores da função proteica que não estão sob uso terapêutico (genericamente chamados de ligantes). Em decorrência de sua variedade e originalidade química, é extremamente difícil ter, de antemão, parâmetros próprios à sua descrição. Assim, é frequente a necessidade de parametrização dos ligantes em estudo, seguindo as características do campo de força em uso.

Embora os quatro campos de força citados possuam parâmetros para um amplo espectro de grupamentos funcionais, para casos específicos ferramentas como o servidor PRODRG (para o GROMOS) e o GAFF (para o AMBER) são capazes de gerar parâmetros, com graus variados de precisão, que podem ser empregados no estudo de compostos orgânicos em geral.

8.8. Conceitos-chave

Amostragem: refere-se à descrição do comportamento conformacional de uma dada molécula em uma simulação.

Campo de força: conjunto de equações que descreve o comportamento molecular em cálculos de mecânica molecular. É ajustado para cada tipo de molécula a ser estudado.

Campo de força *all atom* (todos os átomos): considera todos os átomos do sistema explicitamente.

Campo de força *united atom* (átomo unido): transforma grupos CH, CH_2 e CH_3 em uma única partícula ou pseudoátomo, reduzindo o número de átomos a ser descrito.

Grupos CH de anéis aromáticos são descritos explicitamente.

Campo de força *coarse-grained*: transforma grupos de átomos em partículas, reduzindo o custo computacional ainda mais do que campos de átomo unido.

Condições periódicas de contorno: condição empregada em simulações por DM que impede o contato das moléculas do sistema com o vácuo, representando o sistema de forma periódica.

Cut-off: representa um corte no cálculo de interações não ligadas, reduzindo o custo computacional do cálculo. A partir da distância definida, estas interações não são mais calculadas.

Diedro próprio: ângulo formado por quatro átomos ligados em sequência. Os primeiros três átomos definem um plano, enquanto os últimos três definem outro plano. O ângulo formado por estes dois planos é o diedro.

Diedro impróprio: ângulo formado por quatro átomos que não estão ligados em sequência. É empregado para garantir, por exemplo, a quiralidade de átomos e a planaridade de anéis.

Dinâmica molecular: tipo de cálculo em que as coordenadas dos átomos variam como função do tempo.

Equilibração: período em que propriedades de uma simulação de DM demoram para atingir um patamar estável. Diferentes propriedades podem requerer tempos diferentes para equilibrar.

Mecânica molecular: tipo de cálculo em que o comportamento molecular é descrito a partir das equações da mecânica clássica ou de Newton.

Mecânica quântica: tipo de cálculo em que o comportamento molecular é descrito a partir das equações da mecânica quântica.

Minimização de energia: tipo de cálculo em que a energia do sistema é reduzida através da otimização das posições atômicas.

Modelo de água explícito: modelo no qual as moléculas de água são descritas pela presença física de seus átomos.

Modelo de água implícito: modelo no qual as moléculas de água são descritas sem a presença física de seus átomos.

NPT: condição de simulação na qual o número de partículas, a pressão e a temperatura permanecem constantes.

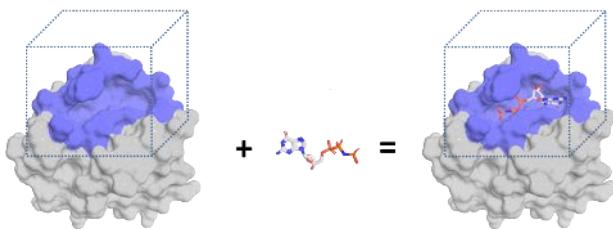
NVT: condição de simulação na qual o número de partículas, o volume e a temperatura permanecem constantes.

Tempo de integração: tamanho do passo empregado em cálculos de DM.

Transferabilidade: em um campo de força, se refere à manutenção das propriedades de um grupamento funcional em diferentes moléculas. Assim, uma hidroxila alcoólica de um resíduo de serina terá os mesmos parâmetros que a mesma hidroxila em uma treonina.

8.9. Leitura recomendada

MORGON, Nelson H.; COUTINHO, K. ***Métodos de Química Teórica e Modelagem Molecular***. São Paulo: Editora Livraria da Física, 2007.

LEACH, Andrew R. ***Molecular Modelling Principles and Applications***. 2.ed. Essex: Pearson Education Limited, 2001.

SANT'ANNA, Carlos Maurício R. Glossário de termos usados no planejamento de farmacos (recomendações da IUPAC para 1997). ***Quim. Nova***, 25, 505-512, 2002.

9. Atracamento Molecular

Isabella A. Guedes
Camila S. de Magalhães
Laurent E. Dardenne

Predição do modo de ligação do GTP no sítio de ligação da proteína c-H-ras p21.

Figura 1-9: Exemplos de complexos moleculares: (A) chaperona GroEL (PDB ID 1AON), (B) complexo DNA com proteína DMT1 (PDB ID 3PT6) e (C) complexo da enzima HIV-1 protease com o inibidor indinavir (PDB ID 1HSG). As versões menores em B e C estão em escala com A.

9.1. Introdução

Para se compreender a maioria dos mecanismos e processos celulares é necessário determinar e compreender o modo de interação entre macromoléculas (principalmente proteínas e ácidos nucleicos) ou entre uma macromolécula e uma pequena molécula ligante, que pode atuar como agonista/antagonista ou substrato/inibidor em determinado processo fisiológico.

Complexos macromoleculares podem envolver dezenas ou centenas de componentes, tais como na formação dos poros nucleares, formação de ribossomos, formação de chaperonas como a GroEL e na formação de capsídeos de vírus (Figura 1-9). Quais proteínas interagem e o modo de interação são informações de fundamental importância para a compreensão do funcionamento de processos biomoleculares.

Por outro lado, o conhecimento do modo de interação entre pequenas moléculas ligantes e proteínas alvo, com um papel crucial em processos fisiopatológicos, é de grande importância para o planejamento racional de fármacos. Neste sentido a técnica computacional denominada atracamento molecular (*molecular docking*, em inglês), dedicada à previsão do modo de ligação e dos detalhes do reconhecimento molecular proteína-proteína e receptor-ligante (Figura 2-9), assume cada vez mais papel de destaque em pesquisa associadas à saúde e à biotecnologia.

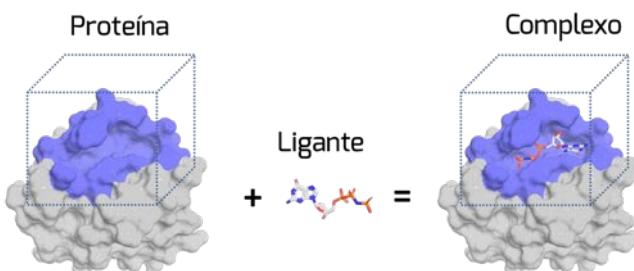



Figura 2-9: Emprego do método de atracamento molecular na predição do modo de ligação do GTP ao seu sítio de ligação na proteína c-H-ras p21.

Os métodos de atracamento molecular envolvem desafios teórico-computacionais formidáveis, e se dividem em duas classes de métodos distintos: receptor-ligante e receptor-proteína. Embora proteínas sejam os receptores mais comuns, outras biomoléculas também podem exercer este papel. Diversos fármacos, por exemplo, modulam diretamente o DNA que, assim, passa a ser o receptor alvo. Adicionalmente, fármacos podem atuar modificando propriedades físico-química da célula, sem necessariamente envolver um processo de atracamento, como na modulação da fluidez de membranas plasmáticas. Neste capítulo, será dada mais ênfase aos métodos de atracamento proteína-ligante, contextualizados dentro da área de planejamento racional de fármacos baseado em estruturas.

9.2. Reconhecimento molecular

As metodologias computacionais de atracamento proteína-ligante estão baseadas no modelo chave-fechadura, proposto por Emil Fischer em 1894. Neste modelo, o receptor proteico é associado à uma "fechadura", e seu sítio de ligação ou sítio receptor é considerado como o "buraco da fechadura". A possível "chave da fechadura" é o ligante, e a interação entre o ligante e a proteína está relacionada a uma das possíveis ações de "abrir ou trancar" a porta.

O modelo chave-fechadura, contudo, induz a uma interpretação de que a "fechadura", representada pela molécula receptora, é rígida. Entretanto, no meio biológico, tanto o ligante quanto a proteína são flexíveis, podendo modificar a sua conformação durante o processo de formação do complexo receptor-ligante. Uma visão mais adequada deste processo é denominada de encaixe induzido, onde tanto o ligante quanto a proteína se adaptam um ao outro durante o processo de reconhecimento molecular (Figura 3-9). De fato, a flexibilidade de uma proteína está diretamente associada à sua atividade, seja na catálise de reações enzimáticas, na transdução de sinais, no transporte através de proteínas de membrana, ou em mudanças conformacionais associadas a formas ativas e não ativas de proteínas.

Uma visão mais moderna do atracamento proteína-ligante descreve uma proteína como um conjunto de

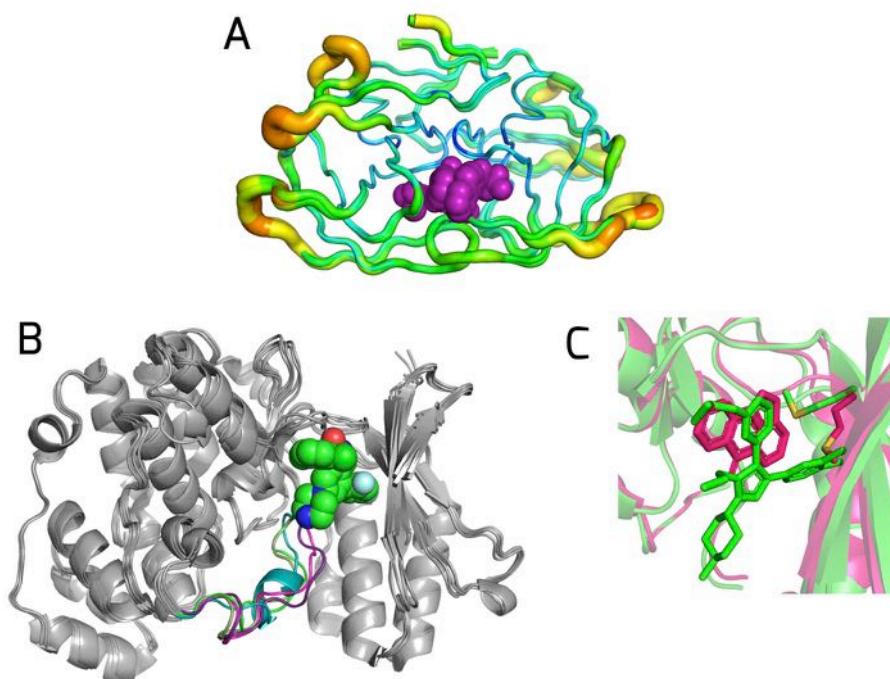



Figura 3-9: Graus de flexibilidade do receptor: (A) mobilidade do esqueleto peptídico da enzima protease do HIV-1, (B) diversas conformações de alça no sítio de ligação do ATP à enzima MAP cinase p38, e (C) mudança conformacional da cadeia lateral de resíduo na enzima cinase JNK3, influenciada por diferentes inibidores.

estados conformacionais, com estruturas similares e energeticamente equivalentes. Nesta visão, ao interagir com determinada proteína, um ligante seleciona uma determinada conformação entre as preexistentes (com a qual possui maior afinidade) e desloca o equilíbrio químico de tal forma que esta conformação tenha a sua proporção aumentada na população total de estados. É importante ressaltar que estudos experimentais sugerem que estes dois mecanismos, encaixe induzido e seleção conformacional, podem coexistir em um mesmo sistema ligante-receptor. Estas visões são muito importantes para direcionar as metodologias de atracamento proteína-ligante no sentido de fornecer um tratamento adequado do problema da flexibilidade intrínseca do receptor proteico.

A introdução da flexibilidade do receptor proteico é um dos maiores desafios das metodologias de atracamento proteína-ligante. Em parte, isto se deve ao fato de que determinadas mudanças conformacionais importantes para a função de proteínas são difíceis de serem caracterizadas experimentalmente e/ou computacionalmente por envolverem milhares de graus de liberdade. Tal complexidade leva estes processos a ocorrerem em escalas de tempo desde microssegundos a vários minutos, envolvendo amplitudes de deslocamento de até dezenas de angstroms (1 Å = 10^{-10} m).

O reconhecimento molecular proteína-ligante está baseado na complementaridade de características físico-químicas e estruturais das moléculas interagentes. As características físico-químicas definem o grau de afinidade e de especificidade do ligante pela proteína, e estão relacionadas com as interações intermoleculares existentes no complexo. Estas interações incluem as ligações de hidrogênio, as interações provenientes do efeito hidrofóbico, as interações de van der Waals, as interações eletrostáticas e as ligações covalentes que possam ser formadas durante o processo de interação receptor-ligante. As características estruturais, por sua vez, estão associadas aos arranjos espaciais moleculares, dados por variações na orientação, posicionamento espacial e rotações de ligações químicas das moléculas interagentes.

Ligantes e proteínas que possuem uma alta afinidade um pelo outro exibem as seguintes características:

i) alto nível de complementaridade estérica, ou seja, a proteína e o ligante possuem uma alta porcentagem de suas superfícies de contato moleculares, definidas pelos raios de van der Waals atômicos, em contato próximo;
ii) alta complementaridade de propriedades associadas às superfícies de contato moleculares (esta complementaridade pode ser tanto eletrostática, onde grupos polares/carregados do ligante ficam perto de grupos da proteína com polaridade/carga complementar, quanto relacionada à complementaridade de regiões hidrofóbicas);
iii) o ligante geralmente se liga em uma conformação energeticamente favorável, e
iv) interações repulsivas entre ligante e proteínas são minimizadas.

Interações proteína-ligante

Os principais tipos de interações intermoleculares envolvidas no reconhecimento molecular proteína-ligante incluem:

i) ligações de hidrogênio;
ii) interações de van der Waals;
iii) interações iônicas;
iv) interações hidrofóbicas;
v) interações do tipo cátion-π;
vi) interações envolvendo anéis aromáticos do tipo π-π e empilhamento-T, e
vii) coordenação com íons metálicos.

O efeito hidrofóbico origina-se do fato de que partes apolares do ligante e do sítio ativo interagem com o solvente, sendo que estas se encontram solvatadas por camadas de moléculas de água mais organizadas. A aproximação destas partes apolares, durante a interação proteína-ligante, liberam e desorganizam as moléculas de água, aumentando a entropia do sistema e consequentemente favorecem a formação do complexo proteína-ligante. O aumento na entropia do solvente associado ao ocultamento das superfícies apolares é chamado de efeito hidrofóbico.

Este efeito destaca o papel fundamental do solvente aquoso no processo de reconhe-

cimento molecular proteína-ligante. Em algumas situações, as moléculas de água assumem tal importância que sua presença é considerada estrutural, sendo por isso denominadas moléculas de água estruturais.

Estas moléculas estão ligadas fortemente ao sítio ativo, e geralmente são conservadas em sítios de ligação de proteínas homólogas. A presença destas moléculas nos sítios receptores de proteínas podem interferir no acesso do ligante ao sítio ativo e modificar o perfil de formação de ligações de hidrogênio, contribuindo portanto diretamente no sucesso das metodologias de atracamento proteína-ligante.

Durante a formação do complexo ocorre a perda de entropia rotacional e translacional do ligante, além de variações na sua entropia vibracional e conformacional devido às restrições de comprimento de ligação, deformação angular e ângulos diedrais. Estas também são contribuições entrópicas importantes que ocorrem durante o processo de reconhecimento molecular.

O processo de reconhecimento molecular proteína-ligante é dirigido por uma combinação de efeitos entálpicos e entrópicos. Estes efeitos podem ser estimados através da energia livre de ligação de Gibbs que, por sua vez, está diretamente relacionada à constante de equilíbrio de ligação K_{eq}, a qual pode ser medida experimentalmente.

$$\Delta G_{lig} = \Delta H - T\Delta S = -RT \ln K_{eq}$$

onde ΔH é a variação de entalpia, T é a temperatura absoluta, ΔS é a variação de entropia e R é a constante universal dos gases.

A constante de equilíbrio de ligação K_{eq} é determinada experimentalmente com relação a um estado de referência (usualmente, para sistemas biológicos, utilizando uma concentração de 1 M e 25 °C). Esta constante de equilíbrio pode ser representada pela constante de dissociação (K_d) ou de associação (K_a), as quais dependem da representação da reação química sendo uma o inverso da outra.

$$K_d = ([R][L])/[RL] \qquad K_a = [RL]/([R][L])$$

onde $[R]$, $[L]$ e $[RL]$ são as concentrações de receptor, do ligante e do complexo receptor-ligante respectivamente.

A determinação destas constantes depende fortemente da temperatura, pressão, pH e força iônica da solução. Para comparar a afinidade de moléculas distintas por um mesmo receptor obtidas por grupos de pesquisa distintos é necessário que os experimentos tenham sido realizados sob as mesmas condições.

Tanto as contribuições entálpicas quanto entrópicas são importantes para a interação receptor-ligante. Muitas vezes, há uma compensação entre estas duas contribuições, podendo a ligação ser determinada principalmente pela contribuição entálpica (compensando uma perda entrópica) ou pela contribuição entrópica (compensando uma variação de entalpia positiva).

A energia livre de ligação de Gibbs pode ser obtida através de métodos teóricos, embora a obtenção de estimativas mais precisas envolva um custo computacional muitas vezes proibitivo para estudos de atracamento molecular em larga escala envolvendo dezenas, centenas ou milhares de ligantes. Alguns dos métodos mais comumente utilizados para cálculo da energia livre incluem o método de perturbação da energia livre (PEL) e o método de integração termodinâmica (IT), que procuram calcular diferenças entre as energias livres de ligação entre ligantes similares.

Embora esses métodos sejam precisos, com erros de aproximadamente 1 kcal/mol, o alto custo computacional envolvido limita a sua utilização. Esses métodos necessitam do conhecimento prévio da estrutura de um complexo onde a proteína está associada com um ligante com estrutura similar ao que se quer estudar. Além disso, tendem a ter um pior desempenho quando os compostos envolvidos diferem de muitos átomos e/ou promovem mudanças conformacionais significativas no receptor. Métodos ainda mais poderosos (conhecidos na literatura como *Absolute Binding Free Energies Methods*), e com custos computacionais mais elevados, procuram calcular os valores das energias livres de ligação sem a necessidade de se ter previamente como referência o conhecimento da energia livre de ligação de um ligante similar.

Uma metodologia mais simples e bastante utilizada para a obtenção de energias livres de ligação é a chamada Energia de

Interação Linear (LIE, do inglês *Linear Interaction Energy*), a qual trata de estimar as energias livres a partir de simulações de dinâmica molecular utilizando um campo de força molecular clássico. Os cálculos de energia livre com esta metodologia envolvem simulações somente nos estados inicial (ligante em solução) e final (complexo receptor/ligante), podendo reduzir desta maneira os problemas de convergência e custo computacionais associados às técnicas PEL e IT. A ideia principal é considerar as contribuições polares e não polares separadamente. A parte polar ou eletrostática pode ser tratada usando a aproximação de resposta linear, enquanto que a não polar é calculada usando uma fórmula empírica calibrada sobre um conjunto de dados experimentais:

$$\Delta G_{\text{lig}}=\alpha(\langle V^{LJ}\rangle_{lig} - \langle V^{LJ}\rangle_{livre}) + \beta(\langle V^{el}\rangle_{lig} - \langle V^{el}\rangle_{livre})$$

onde α é o fator empírico que surge das interações não polares e β é o correspondente às interações eletrostáticas. V representa os valores médios da energia de interação entre o ligante e o meio circundante, tanto para o termo eletrostático (el) como para o de Lennard-Jones (LJ). O método de Energia de Interação Linear tem sido aplicado com sucesso em sistemas complexos, o que o torna um método eficiente e mais rápido para a determinação de energias livres de ligação, mas com um custo computacional suficientemente grande para torná-lo praticamente inviável para estudos envolvendo várias dezenas ou centenas de ligantes.

Outro método utilizado para se obter melhores predições para as energias livres de ligação é o MM-PBSA (*Molecular Mechanics Poisson-Boltzmann Surface Area*) e MM-GBSA (*Molecular Mechanics Generalized-Born Surface Area*). Estes métodos utilizam simulações de dinâmica molecular do ligante/proteína livres e do complexo como base para os cálculos da energia potencial média e de solvatação.

A obtenção de uma descrição suficientemente acurada e viável computacionalmente do papel das moléculas de água no processo de reconhecimento molecular e a quantificação correta das variações entrópicas conformacionais das moléculas interagentes são alguns dos maiores desafios para o desenvolvimento das metodologias de atracamento molecular.

9.3. Métodos de atracamento

O problema de atracamento molecular pode ser dividido em duas partes principais:

i) investigação e predição da conformação e orientação de uma molécula ligante no seu sítio de complexação;

ii) predição da afinidade em um complexo receptor-ligante, isto é, a energia livre de ligação (normalmente chamado na literatura de função *scoring*).

Atualmente existem diversos programas de atracamento molecular disponíveis (Tabela 1-9), distinguindo-se principalmente pelo método de busca e pela função de avaliação de afinidade empregada. Podem ainda diferir quanto à possibilidade de serem utilizados através de portais ou localmente, de utilização gratuita ou paga, na necessidade de registro e na integração com bancos de ligantes e proteínas.

Tabela 1-9: Portais de acesso para alguns programas de atracamento molecular.

Portal	Programa de atracamento
SwissDock	EADock DSS
DockingServer	AutoDock
DockThor Portal	DockThor
1-Click Docking	AutoDock Vina
DOCK Blaster	DOCK
Docking At UTMB	AutoDock Vina
ParDOCK	Método de Monte Carlo
PATCHDOCK	PatchDock
MEDock	MEDock

Preparação do sistema

Uma etapa muito importante para um estudo de reconhecimento molecular proteí-

na-ligante é a preparação do sistema. O primeiro passo nesta etapa é a obtenção das coordenadas das estruturas tridimensionais das moléculas interagentes. Com relação à proteína, o *Protein Data Bank* é atualmente a maior fonte pública de estruturas de proteínas e ácidos nucleicos resolvidos experimentalmente através, principalmente, das técnicas de difração de raios-X e RMN. Na ausência de dados experimentais, estruturas tridimensionais de proteínas podem ser obtidas utilizando-se técnicas de predição de estruturas baseadas em modelagem comparativa ou outros métodos, tais como técnicas baseadas em fragmentos e técnicas baseadas em primeiros princípios.

As estruturas de ligantes podem ser obtidas de vários bancos de dados contendo milhares a milhões de ligantes no formato 1D (smi, *simplified-molecular input-entry system*, também chamado de formato SMILES) ou 2D (sdf, *structure-data file format*, também suporta formato 3D). A geração de uma estrutura 3D de um ligante a partir de uma representação 1D ou 2D (Figura 4-9) pode ser feita através de vários programas tais como, CORINA, CONCORD, OMEGA, Balloon e Multiconf-DOCK.

Uma vez que as estruturas 3D das moléculas tenham sido obtidas, vários cuidados devem ser tomados durante a preparação dos arquivos de entrada para a realização de cálculos de atracamento molecular. Com relação ao sítio de ligação em uma proteína alvo, é necessário primeiramente que se tenha a informação da localização do mesmo. Em um segundo momento, é muito importante realizar um estudo das características físico-químicas e estruturais deste sítio. No caso de enzimas, um estudo (incluindo uma pesquisa bibliográfica) para obter o máximo de informações sobre a reação enzimática envolvida também deve ser realizado.

Como a localização do sítio receptor de uma proteína nem sempre é conhecida, métodos computacionais podem ser utilizados para prever os possíveis sítios de ligação. Estes métodos podem se basear em análises geométricas e de volume para identificar cavidades (tais como FINDSITE, SURFNET e LIGSITE), em energias de interação (Q-SITEFINDER e GRID) e no uso de propriedades de sítios de ligação conhecidos para efetuar uma busca por padrões (webPDBinder).

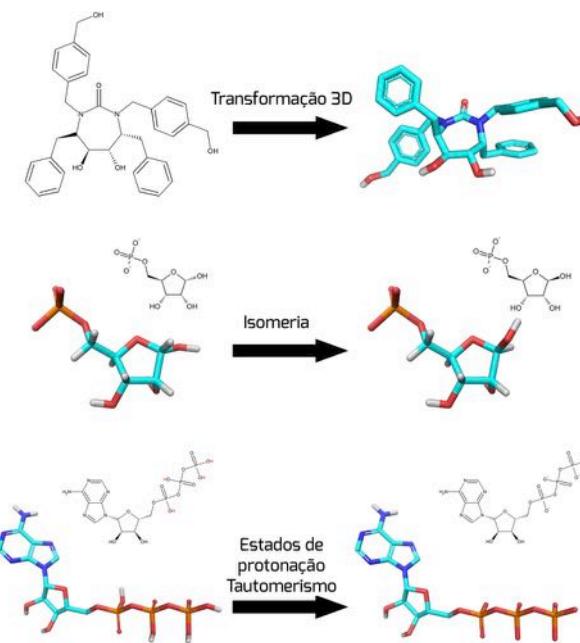



Figura 4-9: Principais etapas de preparação do ligante.

Mesmo quando se tem uma estrutura tridimensional determinada experimentalmente, é importante que se faça uma investigação minuciosa da estrutura na região do sítio ativo à procura de erros (programas como WHAT_IF, MOLPROBITY e PROCHECK podem ser utilizados para checar a qualidade da estrutura e corrigir alguns tipos de erros). Alguns dos possíveis problemas que podem ser encontrados são:

i) ausência de átomos e/ou resíduos;

ii) mal posicionamento de cadeias laterais, particularmente importante para os resíduos de asparagina, glutamina e histidina, onde as cadeias laterais podem apresentar inversões, tais como a inversão entre os átomos OG e ND na asparagina;

iii) presença de duas ou mais conformações para um resíduo ou conjunto de resíduos representando configurações alternativas para a mesma proteína;

iv) conformações não nativas, seja de uma cadeia lateral ou de uma estrutura $2^{\text{ária}}$, devido a efeitos de empacotamen-

to das proteínas no cristal.

Um segundo aspecto de grande relevância na preparação do sítio receptor é estabelecer o estado de protonação correto dos resíduos que participam da interação com o ligante (Figura 5-9). É muito comum que resíduos como cisteína, glutamato, aspartato e histidina tenham estados de protonação não usuais, influenciados e estabilizados pelo ambiente eletrostático do sítio ativo. Este problema pode ser tratado utilizando estratégias complementares, tais como:

i) análise de diferentes complexos (muitas vezes de proteínas homólogas) com distintos ligantes;

ii) estudo da literatura a respeito do mecanismo de reação enzimática;

iii) uso de programas para prever o pKa de cada resíduo do sítio ativo/receptor (por exemplo, através do programa PROPKA).

Com relação ao ligante, a etapa de preparação envolve diversos cuidados, tais como a determinação do seu estado de protonação, estado tautomérico, forma enantiomérica ativa biologicamente (Figura 4-9), a identificação das suas ligações químicas flexíveis (Figura 6-9) e, a partir destas, a geração de múltiplas conformações.

A determinação do estado de protonação do ligante é uma tarefa não trivial, pois envolve não só o pH mas também a interação com o sítio de ligação. Para tentar minimizar este problema, muitas vezes o atracamento é feito levando-se em conta os vários estados de protonação do ligante.

A geração de várias conformações para o ligante é importante no caso de metodologias de atracamento que não levam em conta a flexibilidade do mesmo e fazem o atracamento do ligante rígido para cada conformação representativa. Um caso específico está relacionado a estruturas cíclicas, cuja flexibilidade geralmente não é levada em consideração durante o processo de atracamento.

Ligantes contendo estruturas cíclicas não aromáticas podem exibir mudanças conformacionais relevantes no processo de reconhecimento molecular. Para estes casos, a geração de um conjunto de estruturas representativas das mudanças conformacionais e a utilização destas em múltiplos estudos de atracamento é a solução indicada. LIGPREP é um exemplo de programa que gera tautômeros, diferentes conformações de estruturas cíclicas, diferentes estados de protonação de acordo com o pH e diferentes estereoisômeros para um determinado ligante.

Figura 5-9: Diferentes estados de protonação dos aspartatos catalíticos na estrutura da HIV-1 Protease complexada com o inibidor KNI-272. Estrutura determinada por difração de neutrons.

É importante ressaltar que dificilmente metodologia de busca ou função avaliação é capaz de corrigir ou superar os problemas causados por uma má caracterização do estado de protonação de um ligante ou de resíduos de aminoácidos importantes presentes no sítio de ligação. A correta preparação das estruturas 3D do ligante e da proteína, juntamente com a correta determinação das moléculas de água estruturais, são etapas cruciais para obter sucesso na utilização das metodologias de atracamento receptor-ligante.

Algumas metodologias de atracamento

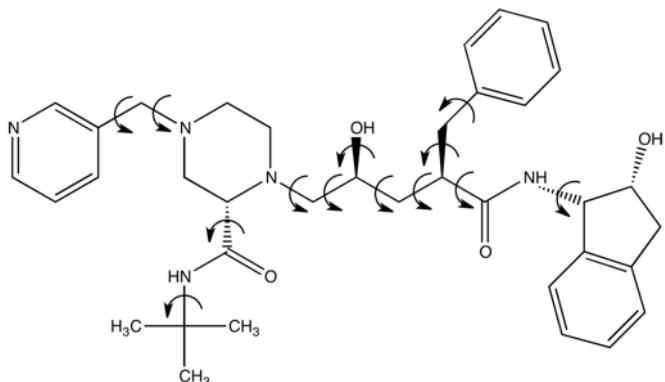



Figura 6-9: Graus de liberdade conformacionais do indinavir, representados por setas.

mais sofisticadas procuram avaliar os diferentes estados de protonação do ligante e das cadeias laterais dos resíduos durante a execução do algoritmo. eHiTS é um exemplo de programa que utiliza este tipo de estratégia.

Métodos de busca

A exploração das diferentes orientações e conformações possíveis para um ligante no sítio de ligação do receptor alvo pelo programa de atracamento deve ser feita de tal forma a se encontrar a solução ótima, ou seja, o mínimo global de energia. Se os efeitos entrópicos e entálpicos associados à termodinâmica do sistema (ou seja, a energia livre do sistema) forem corretamente modelados pela função de energia, então o mínimo global de energia da superfície investigada vai estar associado ao modo de ligação receptor-ligante encontrado experimentalmente. Infelizmente, devido às aproximações introduzidas no modelo de interação molecular, nem sempre o mínimo global satisfaz este importante requisito.

Um ligante pode variar sua orientação dentro do sítio de ligação através de movimentos de translação e rotação (os chamados graus de liberdade translacionais e rotacionais). Além destas modificações, a presença de ângulos diedrais rotacionáveis (isto é, ligações químicas simples) do ligante correspondem aos graus de liberdade conformacionais. Na Figura 6-9 são mostrados os graus de liberdade conformacionais do indinavir, inibidor da protease do HIV-1.

A flexibilidade das moléculas interagentes é considerada de maneira variada pelos diversos métodos de atracamento molecular. Três principais estratégias são utilizadas:

i) a proteína é considerada rígida, e apenas os graus de liberdade translacionais e rotacionais do ligante são considerados, ou seja, o ligante é fixado em uma conformação rígida;

ii) a proteína é considerada rígida, mas todos os graus de liberdade do ligante (translacionais, rotacionais e conformacionais) são levados em conta;

iii) a proteína é considerada totalmente ou parcialmente flexível, e todos os graus de liberdade do ligante também são considerados.

Nas metodologias que utilizam a estratégia *i* é possível considerar a flexibilidade do ligante através da construção prévia de um conjunto de conformações representativas e a subsequente realização de vários cálculos de atracamento molecular do tipo receptor-rígido. De modo análogo, com relação à segunda estratégia, é possível considerar a flexibilidade da proteína em atracamentos do tipo receptor-rígido através da geração de um conjunto de conformações representativo da flexibilidade do receptor proteico.

Os métodos de busca dos programas de atracamento ligante-receptor podem ser classificados basicamente em três categorias: métodos de busca sistemática, métodos de busca determinística e métodos de busca estocástica. Alguns programas utilizam em conjunto algumas destas diferentes abordagens.

Nos métodos de busca sistemática, um conjunto de valores é estabelecido para cada grau de liberdade. O objetivo é explorar de forma combinatória todos os graus de liberdade da molécula durante a busca.

Um dos principais exemplos de métodos de busca sistemática são os algoritmos de construção incremental, um tipo de abordagem baseada em fragmentos. Nestes algoritmos, o ligante é dividido em

pequenos fragmentos rígidos. Em um primeiro momento, um fragmento-base é ancorado no sítio receptor e, posteriormente, todos os outros fragmentos são adicionados de forma incremental, até a reconstrução total do ligante. Cada fragmento adicionado possui uma ligação química rotacionável com o fragmento base. A junção dos fragmentos é feita com base em uma busca conformacional, a partir de um banco de valores de ângulos diedrais, de maneira a investigar sistematicamente a flexibilidade associada a este ângulo específico. Exemplos de programas de atracamento que utilizam construção incremental são DOCK, FlexX, Glide, EUDOC e Surflex.

Nos métodos de busca determinística, dado um mesmo estado inicial de entrada, é obtido sempre o mesmo resultado de saída. Métodos de simulação por dinâmica molecular e métodos clássicos de minimização de energia são exemplos de métodos de busca determinística utilizados por programas de atracamento molecular.

Uma das grandes vantagens dos métodos de atracamento baseados em dinâmica molecular é que tanto a influência do solvente explícito quanto de todos os graus de liberdade do complexo proteína-ligante são explorados de forma mais natural. Entretanto, estes métodos possuem um custo computacional elevado e, dependendo da altura das barreiras de energia encontradas, podem ficar presos em configurações associadas a mínimos locais do sistema.

Para tentar superar esta limitação, é possível utilizar algumas estratégias como, por exemplo, aumentar a temperatura de simulação, suavizar a superfície de energia potencial e simular diferentes partes do sistema proteína-ligante com diferentes temperaturas, além de iniciar os cálculos de dinâmica molecular com o ligante em distintas conformações. O programa CDOCKER é um exemplo de programa que utiliza DM em conjunto com a geração de várias configurações do ligante para serem utilizadas como pontos de partida em simulações com altas temperaturas e potenciais suavizados.

Ainda, uma técnica que tem sido utilizada com bastante sucesso no estudo de interações ligante-receptor é a metadinâmica. Nesta técnica, uma força adicional é calculada durante a simulação de DM. Esta força depende do próprio histórico da simulação, e tem a função de facilitar a amostragem do espaço configuracional do sistema, tentando diminuir a probabilidade de que configurações já visitadas venham a ser amostradas novamente.

Os métodos baseados em DM podem ser utilizados em uma estratégia conjunta com outros tipos de métodos de busca. Nesta estratégia, métodos sistemáticos/incrementais/estocásticos são utilizados para gerar um conjunto de configurações proteína-ligante prováveis. Nesta etapa, muito mais rápida, são introduzidas restrições associadas à flexibilidade do ligante e da proteína, e quanto à descrição do efeito solvente (uso da aproximação de solvente implícito). Na etapa seguinte, muito mais custosa, simulações de DM com solvente explícito e considerando flexibilidade total do receptor e do ligante são realizadas tomando-se como ponto de partida as melhores configurações geradas na etapa anterior.

Nos métodos de busca estocástica o processo de otimização envolve movimentos aleatórios associados aos graus de liberdade. Este fato implica na possibilidade de se obter diferentes resultados como saída para um mesmo estado inicial de entrada. A maioria dos métodos desta classe não possui garantia de convergência. Portanto, em estudos de atracamento molecular, várias execuções independentes do algoritmo são necessárias para se realizar uma boa investigação do sistema. Monte Carlo, Recozimento Simulado (*Simulated Annealing*) e Algoritmos Evolucionistas são exemplos de métodos de busca estocástica mais comumente utilizados por programas de atracamento receptor-ligante. Glide, ICM, Prodock, AutoDock e LigandFit são exemplos de programas que utilizam os métodos estocásticos de Monte Carlo e *Simulated Annealing*.

No método de Monte Carlo padrão (MC) é gerada aleatoriamente uma conformação inicial do ligante e, em seguida, tomando esta configuração como referência, é gerada uma nova configuração. Se a configuração gerada possuir energia menor que a configuração de referência ($\Delta V<0$), a nova configuração é imediatamente aceita e tomada como referência para a próxima iteração. Caso contrário ($\Delta V\geq0$), o critério de Metropólis é utilizado para decidir se a nova configuração será aceita ou não. Esse processo é repetido até que o número desejado de configurações seja obtido.

O critério de Metropólis consiste em se gerar um número aleatório entre 0 e 1 e compará-lo com o fator

de Boltzmann, $\exp(-\Delta V/kBT)$, considerando uma determinada temperatura absoluta T. Se o fator de Boltzman for maior que o número aleatório gerado a nova conformação é aceita. O método de *Simulated Annealing* (SA) pode ser considerado uma variação do método de Monte Carlo, onde o primeiro ciclo da simulação é realizado em uma alta temperatura, sendo que esta decai para temperaturas menores durante os ciclos seguintes. Diferentes variantes de SA utilizam distintas estratégias para o decaimento da temperatura.

O programa MCDOCK utiliza o método SA, o qual também foi utilizado nas primeiras versões do programa Autodock. Prodock e ICM são exemplos de programas de atracamento que utilizam o método de MC com minimização. Neste caso, após um movimento aleatório, a conformação é otimizada por um método baseado em otimização de energia antes que o critério de Metrópolis seja aplicado.

Uma das classes de algoritmos estocásticos mais utilizadas por programas de atracamento molecular proteína-ligante é a de Algoritmos Evolucionistas (AE). Estes algoritmos são inspirados no processo biológico de evolução de populações. Esses algoritmos pertencem à área de Computação Evolucionista (CE), que abrange vários tipos de algoritmos, tais como Algoritmos Genéticos (AG), Estratégias de Evolução (EE), Evolução Diferencial (ED), Otimização por Colônia de Formigas (OCF), Busca Tabu (BT) e Enxame de Partículas (EP). Dentre esses, diversas variantes de Algorimos Genéticos têm sido implementadas para o atracamento de ligantes flexíveis.

AGs são baseados no princípio de sobrevivência do mais adaptado, proposto pela teoria da evolução de Darwin. Ao contrário dos métodos MC e de outros métodos estocásticos que requerem uma única configuração inicial, AGs trabalham com uma população de indivíduos, onde cada indivíduo representa uma possível solução para o problema a ser resolvido. A cada geração, novos indivíduos são gerados através da troca de "genes" entre dois indivíduos "pais" (recombinação) e de mudanças aleatórias nos valores dos "genes" (mutação). Este processo é repetido de maneira que a população evolua para melhores soluções, até que um critério de parada predeterminado seja encontrado.

O primeiro programa de atracamento utilizando AG foi implementado por Judson e colaboradores em 1994, seguido por uma implementação no programa DOCK. O programa de atracamento molecular GOLD utiliza um AG para evoluir múltiplas subpopulações de ligantes, onde a migração entre as populações é permitida. O programa AutoDock também possui implementado um AG convencional e um AG Lamarckiano (AGL). O AGL é um AG híbrido com um método de busca local (BL). A cada geração, uma porcentagem predefinida da população é aleatoriamente escolhida para aplicação da BL. O indivíduo resultante da BL substitui o indivíduo original, em uma alusão à teoria de Lamarck, sobre a hereditariedade de características adquiridas durante o tempo de vida de um indivíduo.

Não há garantia de que os algoritmos evolucionistas encontrem o mínimo global da superfície de energia e, frequentemente, as melhores soluções encontradas ficam presas em mínimos locais. Múltiplas execuções do algoritmo são uma saída óbvia para se tentar uma exploração mais satisfatória do espaço de configurações associado aos modos de atracamento ligante-receptor. Porém, estes problemas tendem a se tornar ainda mais importantes e difíceis de enfrentar quando se lida com ligantes altamente flexíveis (com mais de 10 ligações químicas rotacionáveis) e/ou se considera a flexibilidade da proteína em algum nível.

O programa DockThor (disponível através de portal web www.dockthor.lncc.br) tenta minimizar este problema através do uso de um AG que procura preservar e obter em uma única execução do algoritmo uma multiplicidade de modos de ligação proteína-ligante. Devido à alta complexidade e modalidade (presença de muitos mínimos locais na superfície de energia) desta busca, principalmente para ligantes altamente flexíveis, uma questão crítica é a preservação de diversidade útil na população. O objetivo é permitir a investigação de múltiplas regiões de alta aptidão (nichos) em paralelo, de tal forma a se reduzir as chances de convergência para ótimos locais de baixa qualidade. Para a preservação de múltiplas soluções na população foi proposto o método MRTS (*Modified Restricted Tournament Selection*), baseado no método de seleção por torneio restrito (RTS). O método MRTS possui a vantagem de priorizar a preservação de diversidade

"útil" na população, ou seja, incentiva a preservação de múltiplas soluções de alta aptidão na população ao mesmo tempo em que aumenta a probabilidade de se encontrar o mínimo global.

Os programas MolDock, PRO_LEADS, SODOCK, PSO@Autodock, FIPSDOCK e Autodock Vina são exemplos de programas de atracamento que utilizam estratégias de otimização estocástica. O MolDock utiliza um algoritmo de evolução diferencial. Os programas SODOCK, PSO@Autodock e FIPSDock utilizam variantes do algoritmo de otimização por enxame de partículas (*particle swarm*). O PRO_LEADS utiliza um algoritmo de busca Tabu. O programa AutoDock Vina implementa um algoritmo similar ao utilizado pelo programa de atracamento ICM. Neste algoritmo, uma sucessão de passos consistindo de mutação e busca local são efetuados, onde o resultado de cada passo é aceito ou não de acordo com o critério de Metrópolis.

Funções de avaliação

Os métodos de busca geram uma grande quantidade de conformações do ligante durante o atracamento molecular. As funções de avaliação são combinadas aos métodos de busca para avaliar a qualidade destas conformações de forma a ordená-las de acordo com a sua afinidade pelo receptor. Uma função de avaliação deve ser capaz de distinguir o modo de ligação experimental dos outros encontrados pelo método de busca (ou seja, previsão do modo de ligação). Também deve ser capaz de ordenar corretamente uma lista de ligantes com relação às suas afinidades pela macromolécula receptora (triagem virtual) e prever as respectivas energias livres de ligação (predição de afinidade). Sendo assim, o desempenho de uma função de avaliação está diretamente relacionado à sua capacidade de predição do correto modo de interação do ligante e da sua afinidade pelo receptor alvo.

Estas funções são modelos matemáticos, geralmente lineares, formados por diferentes termos relacionados às propriedades físico-químicas envolvidas na interação de uma pequena molécula ligante com seu sítio de ligação a um receptor. De acordo com o objetivo e a etapa do estudo de atracamento molecular, podem ser utilizadas diferentes funções de avaliação, que variam principalmente no número e tipo de termos, na sua complexidade matemática e na forma de parametrização. Para reduzir o custo computacional, uma função mais simples costuma ser utilizada durante a avaliação das conformações geradas pelo método de busca. Já nas etapas finais do atracamento molecular, uma função de avaliação mais complexa e sofisticada é empregada de forma a obter uma maior acurácia na predição do correto modo de ligação e na predição da afinidade do ligante pelo receptor. As funções de avaliação mais utilizadas no atracamento molecular receptor-ligante podem ser classificadas em três tipos: baseadas em campo de força, empíricas e baseadas em conhecimento.

Funções de avaliação baseadas em campos de força constituem-se em uma soma de termos advindos de algum campo de força molecular clássico, cuja parametrização pode ser feita utilizando dados experimentais ou provenientes de cálculos quânticos (podendo também ser a combinação de ambos). Os termos de energia são divididos em termos não-ligados (associados a interações de van der Waals, eletrostáticas e ligações de hidrogênio) e termos ligados (representando normalmente a energia associada à torção de ligações químicas). Outros termos são normalmente utilizados para tentar incorporar efeitos adicionais, tais como energia de solvatação e interações hidrofóbicas. Exemplos de campos de força moleculares clássicos são GROMOS, AMBER, CHARMM e MMFF94.

As funções empíricas são aquelas desenvolvidas utilizando complexos receptor-ligante com estruturas tridimensionais e afinidades conhecidas. A partir destes dados, seus termos são automaticamente ajustados de forma a reproduzir os dados experimentais de afinidade de ligação com a maior acurácia possível. Neste sentido, estas funções se baseiam na ideia de que a energia livre de ligação pode ser relacionada através do somatório de variáveis não correlacionadas. Cada variável possui um fator relativo de escalonamento, parametrizado de forma a maximizar a correlação com os dados

experimentais. A representação geral de uma função empírica é

$$\Delta G = \sum W_i \cdot \Delta G_i$$

em que W_i é o coeficiente de cada termo ΔG_i referente à determinada propriedade química considerada. A parametrização de uma função empírica tem como objetivo encontrar os valores de W_i que maximizam a correlação da energia de ligação total (ΔG) com os dados experimentais de afinidade de um conjunto de complexos receptor-ligante que treinam o modelo (chamado conjunto de treinamento). Cada função empírica se diferencia no número e nos tipos de termos utilizados, bem como na forma e no conjunto de treinamento utilizado para a sua parametrização. São exemplos de funções empíricas ChemScore, X-Score e GlideScore.

Outro grupo de funções de avaliação são as baseadas em conhecimento. A inspiração para este tipo de função provém da mecânica estatística em sistemas de fluidos simples, que empregam potenciais de força média (*potentials of mean force*, PMF), sendo posteriormente modificadas para serem empregadas em estudos de predição de estruturas de proteínas e estimação de constante de afinidade receptor-ligante.

Estas funções são construídas a partir de análises estatísticas entre os pares de átomos dos complexos receptor-ligante resolvidos experimentalmente. Seus termos são derivados a partir das frequências observadas de interações específicas pré-definidas entre os pares de átomos de cada complexo. Com isto, as funções baseadas em conhecimento tendem a capturar efeitos de interações mais específicas e de modelagem mais complexa. Da mesma forma que as funções empíricas, estas funções se diferenciam pelo tamanho do conjunto de treinamento e no tipo de interações receptor-ligante consideradas durante a parametrização. Uma desvantagem das funções baseadas em conhecimentos é que dependem de um conjunto de treinamento bastante amplo para a parametrização. Além disso, as interações necessárias para construção de uma função baseada em conhecimento podem estar mal representadas no conjunto de treinamento utilizado ou ainda mal parametrizadas, tornando o uso destas funções restrito. Uma vantagem deste tipo de função é que, devido à relativa simplicidade de seus termos, elas conseguem ser tão rápidas quanto as funções empíricas. Alguns exemplos de funções baseadas em conhecimento são DrugScore, RF-Score e PMF.

É importante notar que não existe uma função de avaliação universal, assim como uma classe de função não é necessariamente melhor que outra ou geral o suficiente para ser utilizada com sucesso em qualquer estudo de atracamento. Para obter maior eficiência e confiabilidade, o ideal é utilizar a função de avaliação que mais se adequa ao problema a ser pesquisado. Por exemplo, é necessário saber se todos os tipos de átomos do receptor e do ligante em estudo são definidos na função de avaliação escolhida. Ainda, se a função de avaliação foi parametrizada e testada para a classe do receptor e do ligante estudado. Assim, para estudo de carboidratos, o ideal é utilizar uma função que tenha incluído ligantes desta classe no conjunto de treinamento utilizado na parametrização. Realizar estudos tentando reproduzir complexos determinados experimentalmente (o chamado *redocking*) também auxilia a diagnosticar se a função de avaliação escolhida é capaz de reproduzir os dados experimentais do complexo receptor-ligante (mais frequentemente proteína-ligante).

Estimar a constante de afinidade, como dito anteriormente, ainda é um desafio importante na área da modelagem molecular. Em estudos de triagem virtual, por exemplo, é interessante utilizar mais de uma função de avaliação e comparar os resultados obtidos para chegar a um consenso. Entretanto, a análise qualitativa dos modos de ligação encontrados, tais como a presença de interações intermoleculares consideradas essenciais para o alvo estudado, é de grande importância na detecção de falso-positivos.

Flexibilidade da Proteína

A introdução da flexibilidade da proteína pelos algoritmos de atracamento molecular é atualmente um dos principais desafios desta área de pesquisa. Isto se deve ao grande número de graus de liberdade a serem considerados, principalmente relacionados aos graus de liberdade dos movimentos do esqueleto peptídico e das cadeias laterais dos resíduos de aminoácidos da proteína.

Nos últimos anos, várias metodologias que procuram incorporar este efeito têm sido propostas e descritas na literatura, impulsionadas por dois importantes fatores. O primeiro é que o tratamento da flexibilidade da proteína é cada vez mais reconhecido como um aspecto de extrema relevância em estudos de planejamento racional de fármacos baseado na estrutura do seu receptor biológico. São crescentes as evidências de que alvos moleculares de grande interesse para a indústria farmacêutica passam por importantes mudanças conformacionais quando interagindo com ligantes. O segundo fator foi o grande crescimento do poder de processamento dos computadores ocorrido nos últimos anos, o que tornou possível o desenvolvimento de novas metodologias, algoritmos e abordagens, que seriam inviáveis em estudos de planejamento de fármacos há poucos anos.

A flexibilidade da proteína pode estar associada a diferentes tipos de movimentos, tais como movimentos locais (como o movimento de cadeias laterais de resíduos de aminoácidos localizados no sítio de ligação), movimentos de média escala (como o rearranjo de alças ou reposicionamento de hélices) e movimentos de grande escala, associados a movimentos de domínios da proteína (Figura 3-9). Dependendo dos tipos de movimentos que se quer incorporar, diferentes tipos de metodologias são passíveis de serem utilizadas para um tratamento adequado. De maneira geral, as metodologias existentes podem ser divididas em três categorias, associadas aos três mecanismos de encaixe ligante-proteína mencionados anteriormente:

i) métodos associados ao mecanismo de encaixe induzido, onde são considerados os movimentos locais da proteína;
ii) métodos associados ao mecanismo de conjunto de conformações (*ensemble docking* em inglês), em que são considerados movimentos de grande e larga escala; e
iii) métodos híbridos, que levam os dois tipos de mecanismos e procuram considerar um amplo espectro de movimentos da proteína.

Uma das estratégias mais simples de introduzir a flexibilidade local da proteína é a de suavizar o potencial repulsivo entre átomos do ligante e da proteína, isto é, suavizar o termo de r^{-12} do potencial de Lennard-Jones, técnica esta conhecida na literatura como *Receptor Soft-Docking*. Na prática, isto permite que os ligantes possam se acomodar mais facilmente nas regiões de interação, levando em conta a flexibilidade inerente da proteína. Do ponto de vista da superfície de energia isto corresponde a alargar as regiões de mínimo, evitando assim que um eventual posicionamento incorreto de um átomo da proteína (dentro da aproximação de atracamento com a proteína rígida) possa fazer explodir a energia de interação proteína-ligante, mesmo que esta esteja muito próxima da observada experimentalmente.

Esta técnica também é utilizada para acelerar a convergência da busca conformacional. Normalmente, a intensidade da suavização é utilizada de forma decrescente, permitindo que no início do processo de busca possa haver certa sobreposição entre os átomos do ligante e da proteína. Muitos programas de atracamento utilizam esta suavização embutida na sua função de avaliação. Uma das desvantagens deste método é que ele não é capaz de levar em consideração mudanças conformacionais mais significativas do receptor. Outra desvantagem é a possibilidade de se introduzir erros na avaliação da energia de interação ligante-proteína e de levar muitas vezes à obtenção de falsos positivos e/ou a um conjunto de soluções possíveis cujas energias encontram-se muito próximas, não sendo possível discriminá-las energeticamente.

Os métodos de atracamento mais sofisticados que procuram incorporar a flexibilidade local da proteína simulando um

processo de encaixe induzido fazem isso gerando diversas conformações da proteína concomitantemente com o processo de busca conformacional do ligante dentro do sítio de ligação. Essa abordagem implica em selecionar graus de liberdade adicionais que sejam representativos da flexibilidade da proteína durante o processo de encaixe-induzido. Normalmente, são selecionados graus de liberdade associados a cadeias laterais de resíduos importantes no sítio receptor e, em alguns casos, a regiões específicas do esqueleto peptídico da proteína, tais como alças flexíveis que estejam próximas do sítio e que possam interagir diretamente com os ligantes.

O problema com esta abordagem é que a complexidade do processo de busca cresce a cada grau de liberdade adicionado, aumentando o custo computacional e diminuindo a probabilidade do algoritmo encontrar o mínimo global da superfície de energia. É necessário que o modelador faça uma escolha criteriosa de quais cadeias laterais deve considerar flexíveis. No caso de cadeias laterais de resíduos de aminoácidos, a busca conformacional pode ser feita pela investigação exaustiva dos ângulos torcionáveis da cadeia ou através de uma busca discreta entre conformações preferenciais através da utilização do uso de bibliotecas de rotâmeros. É importante ressaltar que mesmo com a utilização destas bibliotecas, a inclusão da flexibilidade de várias cadeias laterais pode facilmente levar a uma explosão combinatorial que prejudica o desempenho dos algoritmos de atracamento.

Outra estratégia comumente utilizada para introduzir certa acomodação proteína-ligante no processo de atracamento envolve o emprego de um algoritmo de otimização local, tais como aqueles baseados na minimização do gradiente ou em Monte Carlo, para reinvestigar as configurações ligante-proteína geradas durante o processo de busca. O programa Prodock é um exemplo que utiliza a minimização por gradiente durante o processo de busca para incorporar a flexibilidade em regiões da cadeia principal da proteína. O programa ICM/IFREDA utiliza o método de Monte Carlo seguido de minimização de energia para otimizar cadeias laterais e/ou partes flexíveis do esqueleto peptídico. Os programas AutoDock4 e GOLD utilizam algoritmos genéticos para introduzir flexibilidade nas cadeias laterais de resíduos. O programa ROSETTALIGAND utiliza um método de Monte Carlo para explorar simultaneamente os graus de liberdade associados ao ligante, às cadeias laterais dos resíduos e ao esqueleto peptídico da proteína.

Os métodos que se baseiam no mecanismo de conjunto-de-conformações fazem uso de um número discreto de conformações representativas da flexibilidade da proteína ao invés de considerar a flexibilidade da proteína explicitamente durante o processo de atracamento molecular (Figura 7-9). Estas conformações podem ser obtidas de distintos experimentos, utilizando as técnicas de difração de raios-X e/ou RMN. Também podem ser obtidas a partir de modelos gerados por técnicas de predição de estruturas de proteínas, a partir de simulações de dinâmica molecular ou utilizando a técnica de modos normais. Há evidências significativas na literatura de que o uso de múltiplas conformações aumenta significativamente a probabilidade de obter sucesso em estudos de atracamento molecular.

Três questões importantes que se colocam a respeito destas abordagens e que diferenciam os diversos métodos descritos na literatura: *i*) como utilizar as diversas conformações da proteína; *ii*) como gerar e selecionar as conformações da proteína; e *iii*) como ordenar os compostos considerando os atracamentos dos ligantes nas diversas conformações da proteína.

Com relação ao modo de utilização das conformações, a forma mais simples e usual é considerar cada conformação da proteína como rígida e realizar um estudo de atracamento molecular para cada conformação selecionada, embora o custo computacional cresça proporcionalmente ao número de conformações da proteína selecionadas. Uma metodologia de pré-seleção das conformações que reduza significativamente o seu nú-

mero, sem grande perda da informação sobre a flexibilidade do receptor (por exemplo, através de agrupamento por semelhança ou construção de *clusters*), é algo extremamente desejável.

Outra forma possível é o uso de grades de energia (Figura 8-9) combinadas. Os métodos de grade de energia combinada consistem na combinação ou junção de diversas estruturas/conformações rígidas de uma mesma proteína, em uma única grade de energia. A combinação das grades de energia pode ser realizada de várias maneiras. Geralmente, a média ou a média ponderada entre estas grades é calculada, gerando uma única grade. O programa DOCK foi o primeiro a implementar conjuntos de grades de energia para a inclusão da flexibilidade da molécula receptora.

Osterberg e colaboradores compararam vários métodos de grade combinada no programa AutoDock. Um deles utilizava a média entre as grades, outro o valor mínimo e os outros dois utilizavam médias ponderadas. Os resultados obtidos demonstram que a utilização de médias ponderadas é melhor do que a utilização da média e do mínimo. O programa FlexE apresenta um método semelhante, onde a principal diferença reside na forma de tratamento das regiões dissimilares das estruturas do receptor. Os resultados obtidos pelo programa FlexE são de qualidade similar à melhor solução encontrada nos experimentos de atracamento onde cada ligante é atracado em cada uma das conformações representativas da flexibilidade da proteína.

A metodologia de grade é uma estratégia utilizada para aproximar o cálculo das energias eletrostáticas e de van der Waals (outros termos da função energia também podem ser utilizados), reduzindo drasticamente o custo computacional do cálculo da energia de interação intermolecular proteína-ligante. Uma grade de energia pode ser representada como uma malha de pontos tridimensional, em que cada ponto armazena o potencial total eletrostático e de van der Waals. Os valores da energia são obtidos através da interpolação dos valores armazenados nos oito pontos que definem uma célula cúbica da grade. O espaçamento entre os pontos da grade (discretização, `r`) determina o nível da aproximação: quanto maior a discretização, menor a precisão no cálculo da energia de interação intermolecular. O tamanho e formato da grade de energia é dado em função das suas três dimensões (`dX`, `dY` e `dZ`). O centro da grade de energia pode ser definido de diversas formas, como por exemplo centralizar no átomo de um resíduo de aminoácido específico do sítio ativo ou de um ligante de referência. Exemplos de programa que utilizam grade de energia são GOLD, Glide, AutoDock Vina e DockThor.

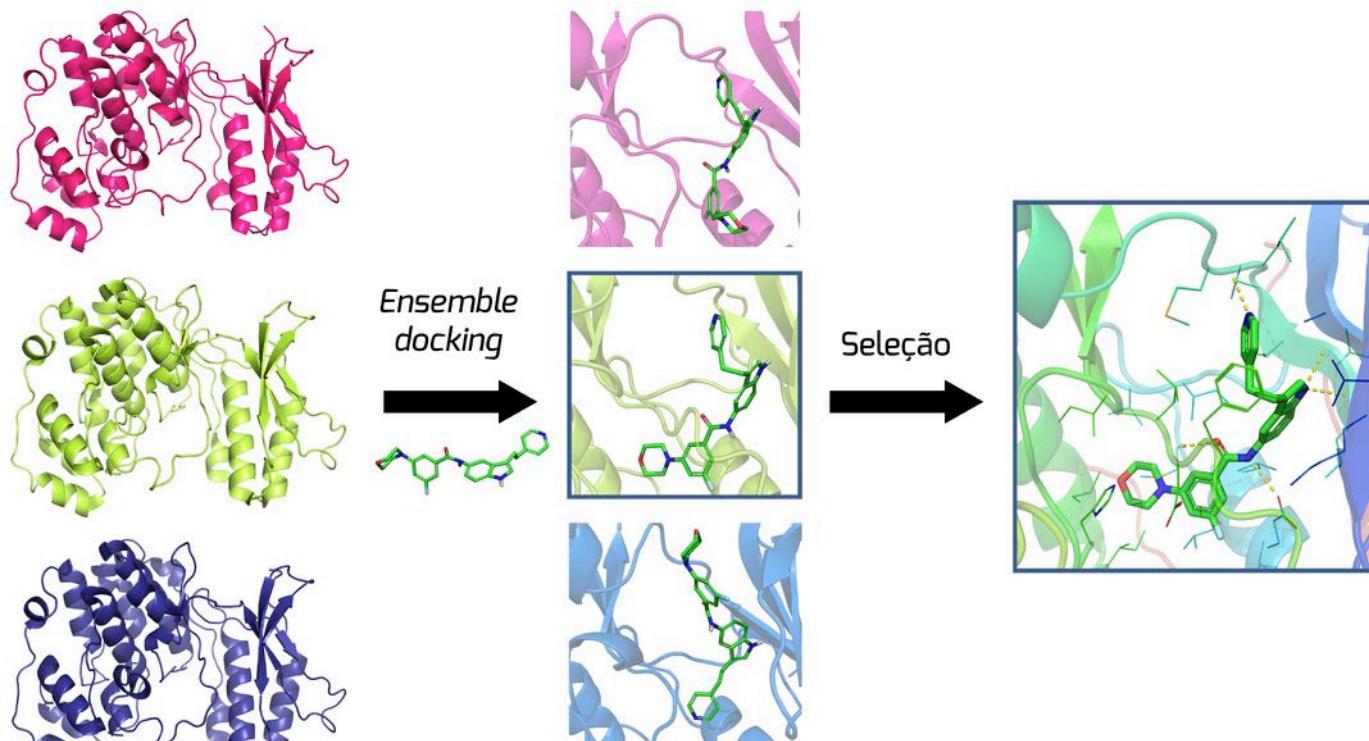



Figura 7-9: Atracamento molecular utilizando conjunto de conformações (adaptado de Guedes e colaboradores, 2013).

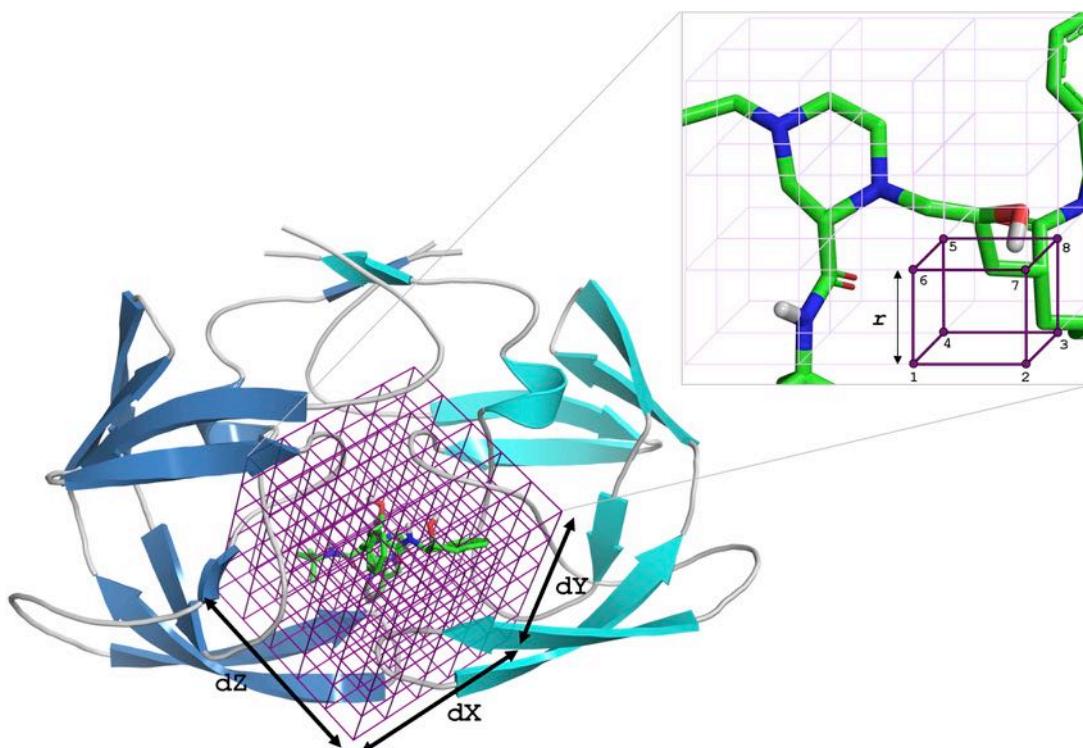



Figura 8-9: Representação de uma grade de energia cúbica centrada no sítio de ligação do inibidor indinavir da protease do HIV-1, com as dimensões de cada eixo (dX, dY e dZ). Em destaque está representada a indexação dos oito pontos de uma célula e a discretização da grade (r). As energias de interação são obtidas da interpolação dos valores, de cada termo da energia, pré-armazenados nos oitos pontos da célula cúbica que contém um determinado átomo do ligante.

Com relação à geração das conformações, as técnicas de simulação de dinâmica molecular e modos normais são as mais utilizadas. Associada ao uso destas técnicas, está a importante questão de qual a amplitude de movimentos do receptor proteico é necessária considerar. Ou seja, se estamos tratando da flexibilidade local de um receptor (como o movimento de uma alça) ou de movimentos de mais larga escala (como movimentos de domínios da proteína). Esta importante questão está diretamente relacionada com a capacidade de amostragem do espaço de configurações do receptor por parte da técnica de simulação utilizada.

Um exemplo de metodologia que usa a técnica de dinâmica molecular é o *Relaxed Complex Scheme*, que utiliza simulações longas de dinâmica molecular considerando todos os átomos do sistema ligante-proteína-solvente. A escala de tempo das simulações variam de 2 ns a 0,5 μs. Uma questão importante a respeito desta técnica é se as simulações devem ser realizadas com a proteína na sua forma apo (não complexada a um ligante) ou na sua forma holo (complexada a um ligante). Resultados descritos na literatura indicam que simulações na forma holo produzem resultados melhores, dando uma descrição mais adequada do sítio de ligação. Na realidade, para não se obter um viés para um determinado modo de ligação de um ligante específico, a estratégia recomendada é a de se realizar várias simulações com ligantes distintos. Estes modos de ligação podem ser obtidos de resultados experimentais ou a partir de resultados obtidos de simulações de atracamento molecular considerando vários ligantes e o receptor rígido.

A questão do número de conformações e de como selecionar aquelas representativas do processo em estudo é ainda uma questão em aberto e possivelmente dependente do tipo de sistema avaliado. Uma das metodologias mais populares busca capturar a diversidade estrutural presente na simulação utilizando o agrupamento de configurações a partir do valor de RMSD (*Root-Mean-Square Deviation*). É importante ressaltar que, neste processo, ao invés de se utilizar a estrutura

de toda a proteína, são normalmente utilizadas as informações relativas a alguns resíduos chave no sítio de ligação da proteína. Normalmente, por questões associadas ao custo computacional, procura-se selecionar um conjunto entre 5-10 conformações.

A questão de como ordenar os compostos levando-se em conta os atracamentos do ligante nas diversas conformações da proteína também não é uma questão fácil de ser respondida. Uma solução é simplesmente utilizar a média das energias dos ligantes com relação às múltiplas conformações da proteína. Outra possibilidade é considerar a melhor/menor energia obtida por um ligante ao interagir com determinada conformação. Existem estudos na literatura que mostram a importância de se considerar ligantes que se ligam fortemente a um conjunto específico (e muitas vezes de baixa probabilidade de ocorrência) de configurações da proteína. São justamente estes casos os mais interessantes, pois abrem oportunidades de desenvolvimento de novos fármacos associados a modos de ligação não usuais.

Outra abordagem utilizada é a reavaliação da energia de ligação utilizando metodologias mais sofisticadas. Um dos grandes problemas com esta técnica é o custo computacional das simulações de dinâmica molecular. Este problema se torna ainda mais importante quando estão envolvidos movimentos de larga escala da proteína. Nestes casos é possível que técnicas como DM acelerada, tais como *Replica Exchange*, metadinâmica e DM utilizando a aproximação para solvente implícito possam ser utilizadas para se obter uma melhor amostragem do espaço das configurações.

O uso das técnicas de Análise de Modos Normais e Análise de Componentes Principais (PCA, *Principal Component Analysis*) para investigar movimentos de larga escala de proteínas talvez sejam as melhores opções para obter uma boa amostragem de conformações em estudos de atracamento envolvendo a técnica de conjunto de conformações.

A técnica de Análise de Modos Normais procura caracterizar os modos de vibração de baixa frequência, os quais se espera estarem associados aos movimentos funcionais de larga escala da proteína. A partir da diagonalização da matriz Hessiana, obtida das derivadas segundas da função energia potencial associada a um campo de força clássico, obtém-se as direções de movimento dos átomos (associadas aos autovetores da matriz) e as frequências de vibração (associadas aos respectivos autovalores). Versões mais simplificadas da técnica de modos normais têm sido desenvolvidas nos sentido de permitir o uso da técnica em sistemas muito grandes. O método conhecido como *Elastic Normal Mode* simplifica o sistema molecular de tal modo que apenas os carbonos alfa da proteína, conectados por potenciais harmônicos, sejam considerados.

Já a técnica PCA utiliza as configurações geradas por uma DM para identificar os graus de liberdade coletivos da proteína. Esta técnica também implica na diagonalização de uma matriz, nesta caso, a matriz de correlação dos movimentos dos átomos da proteína, sendo que os autovetores associados aos maiores autovalores se referem aos movimentos de mais larga escala.

Dependendo do sistema em estudo é desejável que seja feita uma combinação das técnicas anteriormente descritas. Neste sentido, conformações geradas utilizando a técnica de Modos Normais para refletir movimentos amplos da proteína podem servir de base para estudos de DM relativamente curtas. Estas irão refletir o arranjo local das cadeias laterais associado àquela região do espaço de configurações.

Estas configurações utilizadas no contexto da técnica de conjunto de conformações podem ser investigadas com métodos de atracamento baseados no mecanismo de encaixe induzido ou em uma abordagem utilizando grades de energia combinada.

9.4. Triagem em larga escala

Cada vez mais as indústrias farmacêuticas e os grupos de pesquisa que trabalham na busca de moléculas candidatas a novos fármacos necessitam de metodologias mais rápidas, eficazes e de baixo custo. Neste cenário, a triagem virtual (*virtual screening*, em inglês) tem se destacado como uma importante ferramenta na busca de compostos

promissores. A triagem virtual consiste em analisar computacionalmente uma grande quantidade de ligantes com o objetivo de selecionar, de acordo com algum critério predefinido, compostos provavelmente mais ativos frente a determinado alvo farmacológico (ou seja, um receptor). Esta abordagem pode ser empregada para complementar os resultados obtidos pela triagem experimental (*high-throughput screening*, em inglês).

A busca dos ligantes para o estudo de triagem virtual pode ser feita em bancos de estruturas de compostos disponíveis através de portais *online*, tais como ZINC, BindingDB, PubChem, SuperNatural e ChEMBL. Nestes bancos, a busca pode ser feita utilizando propriedades físico-químicas definidas pelo usuário, como número de ligações rotacionáveis e logP ou, em alguns deles, desenhar o fragmento desejável na estrutura dos ligantes. Estes filtros são comumente utilizados com o objetivo de reduzir o número de compostos a serem analisados pela triagem virtual, especificando o perfil desejado para estes ligantes. Após selecionar a lista de ligantes para serem extraídos, geralmente o banco fornece uma tabela com as principais propriedades químicas dos compostos. Caso seja necessário, como no caso da construção de uma biblioteca de ligantes própria do usuário, é possível usar programas que filtram e quantificam tais propriedades, como o FAF-Drugs.

A triagem virtual pode ser feita utilizando diversas metodologias que, de forma geral, agrupam-se naquelas baseadas na estrutura do receptor (*structure-based*) e naquelas baseadas na estrutura do ligante (*ligand-based*). O método baseado na estrutura é mais utilizado quando a estrutura tridimensional da molécula receptora está disponível com boa qualidade. Nesta metodologia, é realizado um estudo de atracamento molecular de todos os ligantes previamente selecionados, ao invés de apenas uma molécula. É possível, assim como no estudo de atracamento molecular tradicional, considerar a flexibilidade do receptor diretamente pelo programa de atracamento ou utilizar um conjunto de conformações da molécula receptora (*ensemble docking*). Entretanto, o custo computacional aumenta significativamente ao se incluir a flexibilidade do receptor em estudos de triagem virtual.

Quando não é possível obter a estrutura tridimensional do receptor, ainda que por técnicas sofisticadas de predição de estruturas de macromoléculas, então o método baseado na estrutura do ligante é empregado. Esta abordagem consiste na análise de similaridade de propriedades estruturais e físico-químicas de compostos ativos e inativos. Duas abordagens importantes incluem o estudo da relação estrutura-atividade (SAR, *structure–activity relationship* ou QSAR, *quantitative structure–activity relationship*) e a modelagem farmacofórica.

Apesar de a triagem virtual baseada em estrutura ser uma técnica amplamente utilizada, o protocolo escolhido pelo pesquisador necessita ser validado para aumentar a confiabilidade dos resultados. Primeiramente, é preciso avaliar se o método de busca e a função de avaliação escolhidos são capazes de reproduzir o modo de ligação experimental de compostos originalmente complexados com o receptor alvo.

Outra análise que deve ser feita é a capacidade de o protocolo diferenciar as moléculas ativas das inativas, conhecidas como casos falso-positivos. Esta validação é de grande importância na triagem virtual, uma vez que auxilia a reduzir o número de moléculas inativas, limitando assim o número de falsos-positivos.

O cálculo da proporção de moléculas ativas frente ao número de inativas presentes em um conjunto de ligantes com dados de atividade experimental previamente conhecidos pode ser feito pelo fator de enriquecimento (*Enrichment Factor*, EF). As moléculas presumidamente inativas (*decoys*) possuem propriedades físicas similares (tais como massa molecular, número de ligações rotacionáveis, logP, número de aceptores/doadores de ligações de hidrogênio) às ativas, entretanto distintas topologicamente (ou seja, exibem diferentes estruturas químicas). Para validar a função de avaliação, utiliza-se um conjunto de ligantes formado por essas moléculas inativas e por um núme-

ro geralmente pequeno de compostos ativos conhecidos. O estudo de atracamento molecular é realizado, e então o EF é usado para medir a capacidade da função ordenar, nas primeiras posições, determinada fração de compostos ativos frente aos inativos.

O desempenho dos diferentes protocolos de atracamento molecular varia significativamente entre os estudos de validação realizados, sendo influenciado diretamente pela metodologia empregada bem como pela composição do conjunto de de dados utilizado (classe dos receptores e perfil dos ligantes incluídos). Quando o número de compostos ativos e inativos é similar, o método AUC (*area under the receiver operating characteristic*) é mais apropriado para avaliar o desempenho do protocolo de triagem virtual.

Os compostos selecionados, conhecidos como *hits*, são encaminhados para as etapas de síntese química (no caso de compostos apenas planejados ou não disponíveis para compra) e estudos de atividade farmacológica (testes *in vitro* e *in vivo*).

9.5. Considerações finais

A descoberta e planejamento de novos fármacos é um processo muito caro e muito demorado. Para levar um novo fármaco ao mercado são necessários de 10 a 20 anos e o custo estimado é de cerca de 800 milhões de dólares. Abordagens *in silico* que possam reduzir estes custos e acelerar o processo de descoberta e planejamento de novos fármacos são extremamente bem vindas e necessárias. É importante ressaltar que já existem diversos exemplos de moléculas que foram descobertas/otimizadas utilizando técnicas computacionais e que estão na fase de ensaios clínicos ou que já foram aprovadas para uso terapêutico.

É possível prever que, no futuro, metodologias computacionais mais sofisticadas terão um papel cada vez mais destacado em estratégias de planejamento racional de fármacos. Neste sentido, alguns aspectos associados às metodologias de atracamento molecular discutidas neste capítulo necessitam de avanços teórico/metodológicos para que se consiga obter uma melhor previsão das constantes de afinidade receptor-ligante. Alguns destes aspectos são a consideração da rugosidade e forma da superfície de energia associada ao complexo receptor-ligante, a estimativa das entropias associadas ao processo de ligação, a consideração não só de múltiplas conformações (flexibilidade) do receptor mas também de múltiplos modos de ligação do ligante, a consideração das mudanças na estruturação das moléculas de água no sítio receptor e da solvatação/dessolvatação do ligante e a consideração de efeitos de mudança de estados de protonação de resíduos do sítio receptor durante o processo atracamento ligante-receptor.

9.6. Conceitos-chave

Algoritmo: conjunto ordenado de instruções para resolver determinado problema.

Atracamento: método para prever o modo de ligação e a afinidade de ligação de uma macromolécula receptora com outra molécula ligante (seja uma outra macromolécula ou uma molécula ligante pequena).

Desenho racional de fármacos baseado em estrutura: área de pesquisa que abrange os métodos computacionais que utilizam informações da estrutura tridimensional da molécula receptora para descoberta e/ou desenvolvimento de novos fármacos.

Encaixe induzido: modelo que sugere a existência de mudanças conformacionais na molécula receptora e no ligante devido à formação do complexo receptor-ligante.

Função de avaliação: função de pontuação que tem por objetivo quantificar a qualidade das soluções obtidas no atracamento molecular.

Ligante: molécula que interage no sítio de ligação de uma macromolécula para formar um complexo, podendo induzir ou bloquear determinada resposta biológica.

Método de busca: algoritmo utilizado pelo atra-

camento molecular para encontrar os modos de ligação do ligante no sítio receptor. Explora os graus de liberdade translacionais, rotacionais e conformacionais.

pKa: logaritmo negativo da constante de acidez ou constante de dissociação ácida (pK_a = -logK_a). Mede a força de um ácido em solução.

Receptor: macromolécula que possui um sítio de ligação de interesse.

Reconhecimento molecular: mecanismo pelo qual uma molécula se liga a outra com perfil complementar, formando um complexo.

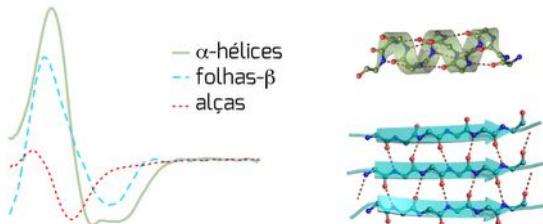
Triagem virtual: metodologia de atracamento molecular em larga escala, através da qual dezenas, centenas ou milhares de ligantes são avaliados no sítio de ligação de um receptor.

9.7. Leitura recomendada

KITCHEN, Douglas B.; et al. Docking and scoring in virtual screening for drug discovery: methods and applications. ***Nat. Rev. Drug Discov.***, 3, 935–949, 2004.

MOBLEY, David L.; DILL, Ken A. Binding of Small-Molecule Ligands to Proteins: ‘What You See’ Is Not Always ‘What You Get’. ***Structure***, 17, 489–498, 2009.

GUEDES, Isabela A.; MAGALHÃES, Camila S.; DARDENNE, Laurent E. Receptor–ligand molecular docking. ***Biophys. Rev.***, 2013.

BROOIJMANS, Natasja; KUNTZ, Irwin D. Molecular recognition and docking algorithms. ***Annu. Rev. Biophys. Biomol. Struct.***, 32, 335–373, 2003.

SPERANDIO, Olivier; et al. Receptor-based computational screening of compound databases: the main docking-scoring engines. ***Curr. Protein Pept. Sci***., 7, 369–393, 2006.

TAYLOR, R. D.; JEWSBURY, P. J.; ESSEX, J. W. A review of protein-small molecule docking methods. ***J. Comput. Aided Mol. Des.*** 16, 151–166, 2002.

TALELE, T. T.; KHEDKAR, S. A.; RIGBY, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. ***Curr. Top. Med. Chem.*** 10, 127–141, 2010.

10. Dicroismo Circular

Representação das curvas de CD associadas a hélices α e folhas β.

Marcelo A. Lima
Edwin A. Yates
Ivarne L. S. Tersariol
Helena B. Nader

10.1. Introdução

O dicroismo circular (CD) é uma técnica espectroscópica utilizada para estudar uma grande variedade de moléculas quirais, tais como fármacos, polímeros e biopolímeros, em solução. Particularmente no caso das proteínas o CD, juntamente à cristalografia de raios-X (capítulo 13), o RMN (capítulo 12), o infravermelho (capítulo 11) e métodos como a modelagem comparativa (capítulo 7) e a dinâmica molecular (capítulo 8), exerce importante papel na busca pelo conhecimento da estrutura e função nucleicas. Tais informações, por sua vez, são essenciais na busca por novos compostos com potencial terapêutico.

Para sistemas enovelados e estruturados tridimensionalmente, como enzimas e proteínas globulares, o CD é uma técnica de baixa resolução quando comparado à RMN e cristalografia de raios-X. Isto ocorre porque o CD, ao contrário destes métodos, não possui resolução atomística, ou seja, não é capaz de identificar átomos específicos das moléculas em estudo.

No entanto, enquanto estruturas desordenadas (ou seja, desenoveladas, forma adotada por aproximadamente a metade das proteínas de mamíferos) tornam-se em grande medida impróprias para estudos de RMN e cristalografia de raios-X, o CD ainda é capaz de lidar com suas estruturas. Além disso, estudos de CD podem ser realizados em solução, em condições bem próximas das fisiológicas, fazendo deste método uma ferramenta ideal para investigar as interações entre moléculas envolvidas nos mais diversos processos biológicos.

Por definição, espectroscopia nada mais é do que o levantamento de dados físico-químicos de um determinado sistema através da transmissão, absorção ou reflexão da energia radiante incidente. No caso do CD, a energia incidente é a ultravioleta comumente na faixa do UV próximo, 380 a 200 nm. Assim, o espectro de CD é gerado pela diferença na capacidade de absorção dos componentes esquerdo e direito da luz circularmente polarizada (mais detalhes adiante) por moléculas quirais que possuem átomos de carbono assimétricos e, consequentemente, diferentes atividades ópticas.

Esta capacidade de absorção de moléculas quirais está diretamente ligada às diferenças nos seus coeficientes de absorbância. Assim, diferentes moléculas ou partes delas possuem CD em regiões específicas do espectro.

Em instrumentos de laboratório, espectros de CD são normalmente registados no

ultravioleta (UV), tipicamente em comprimentos de onda variando de 180 a 260 nm. Além desta região, várias fontes de radiação síncrotron estão disponíveis e possibilitam a obtenção de espectros de CD com intervalos de comprimento de onda consideravelmente maiores. Luz síncrotron é a radiação eletromagnética produzida por elétrons de alta energia através de um acelerador de partículas. Essa luz abrange uma ampla faixa do espectro eletromagnético, incluindo os raios-X, luz ultravioleta e infravermelha, além da luz visível.

De maneira geral, os espectros de CD podem ser utilizados para diversos tipos de estudos, incluindo-se: 1) enovelamento e estrutura 2ária de proteínas; 2) estrutura de proteínas de membrana inseridas em bicamadas lipídicas; 3) interação entre moléculas; 4) interações entre macromoléculas, destacadamente proteínas, ácidos nucleicos e carboidratos; 5) monitoramento da integridade estrutural de moléculas sob aquecimento; 6) quantificação de alterações conformacionais; 7) caracterização de domínios de proteínas, a qual pode ser empregada em comparações com modelos gerados computacionalmente; 8) análise de carboidratos; 9) cinética rápida de enovelamento de proteínas e montagem de complexos macromoleculares, dentre outros.

Além do CD convencional (também chamado de eletrônico, aquele que ocorre na faixa do UV), também existem fenômenos de dicroísmo circular que ocorrem na região do infravermelho, sendo este tipo de fenômeno chamado de dicroismo circular vibracional (VCD). Ele ocorre normalmente entre 3300 e 800 cm^{-1}, e uma de suas principais vantagens em relação ao CD é que, embora as transições eletrônicas tenham uma pequena diferença entre o estado fundamental e o nível excitado, nas transições vibracionais esta diferença é bem maior do que nos espectros contínuos, que possuem sinais distribuídos continuamente em uma certa faixa espectral. Assim, sinais com valores (comprimento de onda) distintos são observados.

O benefício experimental do VCD é que ligantes, como alguns carboidratos, possuem um sinal de CD muito menor quando comparado aos provenientes de uma proteína. Assim, o VCD pode ser utilizado para monitorizar a interação de proteínas com açúcares diretamente e sem a necessidade de manipulação matemática dos espectros.

10.2. Luz polarizada

Para o estudo do CD, um importante conceito que devemos ter em mente é o da luz polarizada. A luz convencional, como a luz solar e a luz de lâmpadas residenciais, são exemplos de luz não polarizada, já que elas emitem radiação que se propaga em todos os planos. Isso ocorre porque a luz branca é composta por ondas eletromagnéticas que vibram em diversos planos perpendiculares à direção da propagação da luz (Figura 1A-10). Por outro lado, a luz polarizada é aquela que possui vibração em apenas um plano (Figura 1B-10).

No caso do CD, a luz utilizada é circularmente polarizada (Figura 2-10), o que nada mais é do que a combinação de duas ondas linearmente polarizadas, uma vertical e outra horizontal, de mesma amplitude.

A diferença de absorção da luz circularmente polarizada à direita e à esquerda dá origem ao espectro de CD. Assim, temos que $CD = AD - AE$, onde AD representa a absorção da luz circularmente polariza à direita e AE a absorção da luz circularmente polariza à esquerda.

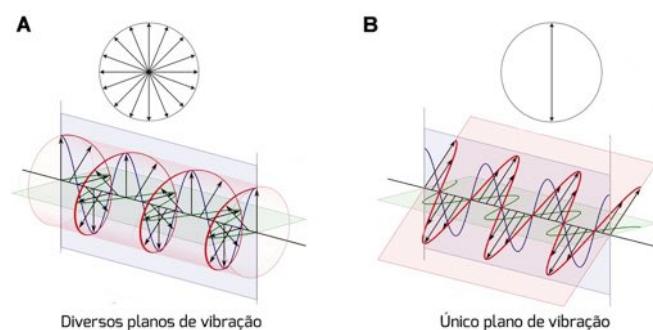



Figura 1-10: Representação planar da luz não polarizada (A) e polarizada (B).

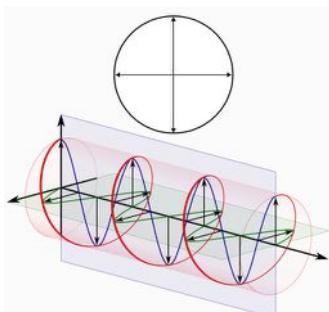



Figura 2-10: Representação planar da luz circularmente polarizada.

10.3. Quiralidade

A quiralidade significa a não sobreposição de sua própria imagem com aquela projetada em um espelho ou, em outras palvras, são imagens que não admitem plano de simetria. Um exemplo clássico de quiralidade é a nossa mão: se colocarmos uma delas diante de um espelho, ela produzirá uma imagem diferente dela própria. A imagem gerada da mão direita será a da mão esquerda e vice-versa. Contudo, as mãos não são sobreponíveis, ou seja, quando sobrepostas não se tornam equivalentes (Figura 3-10). Esta característica é apresentadas por algumas moléculas, que são chamadas assim de isômeros ópticos ou enantiômeros (ver capítulo 2).

No CD, quando a luz polarizada passa através de uma substância quiral, seus componentes podem ser resolvidos e absorvidos com intensidades diferentes. A diferença da absorbância, ΔA, entre a luz polarizada para a direita e para a esquerda, $\Delta A = AD - AE$, está relacionada com seus respectivos coeficientes de absorbância, $\Delta\varepsilon = \varepsilon D - \varepsilon E$, onde εD *e* εE são os coeficientes molares de adsorção da luz circularmente polarizada à direita e à esquerda, respectivamente.

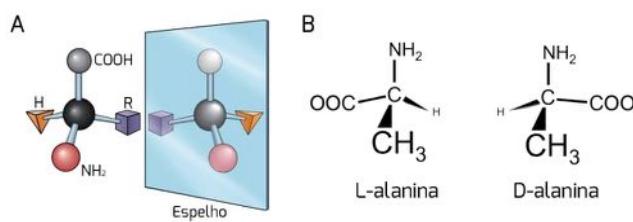



Figura 3-10: Representação da imagem especular (A) de dois enantiômeros do aminoácido alanina (B).

Adicionalmente, sabemos pela lei de Lambert-Beer que $\Delta A = \Delta\varepsilon c l$, onde c representa a concentração da amostra e l o comprimento do percurso óptico. Assim, a resultante de todas essas características darão origem ao espectro de CD de uma dada molécula.

10.4. Instrumentação

Um espectofotômetro de CD pode ser esquematizado segundo apresentado na Figura 4-10. A luz da fonte (L) é dispersa no monocromador (MC), produzindo uma banda estreita de comprimentos de onda que passa através de um polarizador linear (PL).

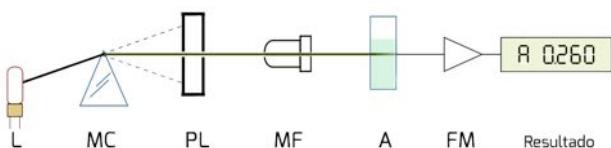



Figura 4-10: Representação esquemática de um espectrofotômetro de CD. Fonte de luz (L); Monocromador (MC); Polarizador linear (PL); Modulador fotoelástico (MF); Amostra (A); Fotomultiplicador (FM). Figura adaptada da Internet.

O polarizador divide o feixe monocromático não polarizado em dois feixes linearmente polarizados. Assim, um dos dois feixes linearmente polarizado passa pelo modulador fotoelástico (MF), que consiste de uma placa transparente e opticamente isotrópica, ou seja, de mesmo índice de refração, ligada a um cristal de quartzo. Quando um campo elétrico alternado é aplicado, a luz que emerge a partir dos interruptores do MF volta com a frequência do campo elétrico aplicado.

Se a amostra (A) possui sinal de CD, a quantidade de luz absorvida varia periodicamente com a polarização da luz incidente e, portanto, a intensidade de luz que atinge o fotomultiplicador (FM) apresenta variações de intensidade sinusoidal na frequência do campo aplicado ao MF. Portanto, o sinal de saída do fotomultiplicador é constituído por um sinal de corrente elétrica alternada sobreposto

a um sinal de corrente elétrica contínua.

Posteriormente, o componente de corrente alternada é filtrado e amplificado. A relação entre a corrente alternada e o componente de corrente contínua é diretamente proporcional ao dicroísmo circular da amostra, sendo esta relação registada em função do comprimento de onda.

10.5. Aplicações a biomoléculas

Proteínas

Na faixa do UV distante, os sinais (ou bandas) relacionadas à ligação peptídica dominam o espectro de CD de proteínas. Este cromóforo apresenta duas transições eletrônicas na faixa do UV distante:

i) transições $n \rightarrow \pi^*$, por volta de 220 nm;

ii) transições $\pi \rightarrow \pi^*$, por volta de 190 nm para amidas secundárias (ligação peptídica para todos os aminoácidos, exceto a prolina), e em torno de 200 nm para amidas terciárias (ligação peptídica envolvendo prolina).

A transição $n \rightarrow \pi^*$ possui coeficiente de absorção fraco, embora dê origem a bandas fortes de CD. Já a transição $\pi \rightarrow \pi^*$ está associada à elevada absorbância e fortes bandas de CD. Devido ao forte momento dipolar de transição eletrônica, as transições $\pi \rightarrow \pi^*$ em ligações peptídicas vizinhas interagem umas com as outras, dando origem a duas ou mais bandas de CD.

As cadeias laterais aromáticas dos resíduos de fenilalanina, tirosina e triptofano possuem fortes bandas de absorbância no UV distante, contribuindo para o espectro de CD de proteínas. Na maioria dos casos, tal contribuição é pequena em comparação com as dos aminoácidos mais numerosos. Porém, para algumas proteínas, as faixas do CD aromático são claramente discerníveis.

No UV próximo, o espectro de CD de proteínas é dominado pelas transições eletrônicas dos grupos aromáticos e ligações dissulfeto. As bandas das cadeias laterais aromáticas são relativamente bem definidas, e possuem uma estrutura característica devido a efeitos vibracionais. Em proteínas com um pequeno número de cadeias laterais aromáticas, as bandas são frequentemente atribuídas a um dos três tipos de resíduos aromáticos e, em alguns casos, através de mutagênese sítio dirigida, a resíduos específicos da sequência proteica. A histidina, apesar de ser um aminoácido aromático, possui um grupamento imidazólico que apresenta sinal de CD abaixo de 220nm e que, em grandes concentrações pode até atrapalhar as medições.

As faixas de CD das ligações dissulfeto são normalmente distinguíveis das faixas de CD aromáticas, já que são menos definidas. Em proteínas que não possuem aminoácidos aromáticos, não há bandas de CD em comprimentos de onda acima de 300 nm. Muitos grupos prostéticos, coenzimas, íons de metais de transição e outros ligantes apresentam bandas de absorbância nesta faixa de comprimento de onda, e estas estão associadas a bandas de CD em complexos com proteínas.

Os diferentes tipos de estrutura 2ária de proteínas (ver capítulo 2) possuem espectros de CD característicos, estabelecidos a partir de modelos de oligo- e polipeptídios com estrutura 2ária conhecida. A Figura 5-10 apresenta os espectros de CD de hélices α, folhas β e estruturas irregulares (desordenadas).

Hélices α apresentam o espectro de CD mais distinto e mais forte, com duas bandas negativas de grandeza comparável por volta de 222 e 208 nm, além de uma forte banda positiva com sua máxima em torno de 190 nm.

A banda em torno de 222 nm resulta das transições $n \rightarrow \pi^*$ do grupo amida, enquanto que as bandas por volta de 208 e 190 nm surgem das transições $\pi \rightarrow \pi^*$ do mesmo grupo. Estas transições $\pi \rightarrow \pi^*$ estão relacionadas a grupos amida mantidos em uma geometria helicoidal bem definida.

As interações entre os momentos dipolares de transição em um arranjo helicoidal dão origem às três bandas de absorbância, uma a 208 nm, polarizada paralelamente ao eixo da hélice, e duas bandas a 190 nm, pola-

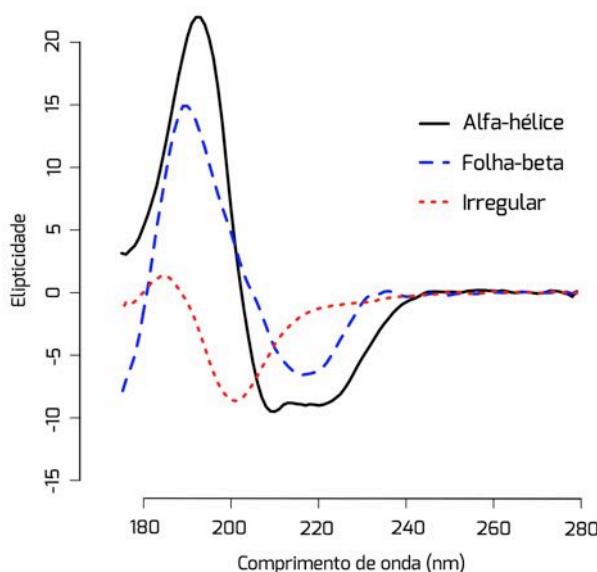



Figura 5-10: Espectros de CD de estruturas do tipo α-hélices, folhas-β e estruturas irregulares.

rizadas em duas direções perpendiculares ao eixo da hélice. Para a hélice à direita, a banda paralela está associada a uma banda de CD negativa a 208 nm, e as bandas perpendiculares com a uma banda positiva a 190 nm.

O CD de uma hélice α é, em sua maioria, independente do solvente e da sequência de aminoácidos. Resíduos aromáticos (Phe, Tyr e Trp) podem modificar o espectro de CD de uma hélice α, especialmente se eles constituem uma fração considerável dos resíduos da proteína. Em homopolímeros de aminoácidos aromáticos, o espectro de CD de uma hélice α é tão distinto que se torna irreconhecível.

O CD de folhas β é bem distinto daquele observado para hélices α, apresentando apenas uma banda negativa de máxima absorbância em 217 nm e uma banda positiva na região entre 195-200 nm como características (Figura 5-10).

O valor absoluto da razão entre a elipticidade do máximo positivo a 197 nm e o máximo negativo a 217 nm amplia-se com o aumento de torção da folha, e é maior para folhas paralelas do que para as folhas anti-paralelas torcidas.

Todos os modelos de polipeptídios com estruturas irregulares (desordenadas) possuem uma forte banda negativa por volta de 200 nm (Figura 5-10). Porém, alguns possuem uma banda positiva em comprimentos de onda maiores e outras um ombro negativo também em comprimentos de onda maiores.

Carboidratos

O CD tem aplicações importantes no estudo de carboidratos, embora estes sejam mais limitadas do que para as proteínas e ácidos nucleicos. Dos cromóforos comuns aos carboidratos, apenas o grupo amida (açúcares N-acetilados) e grupos carboxila (ácidos urônicos) possuem bandas de CD acima de 200 nm. Grupamentos éter, hidroxila, acetal e cetal apresentam suas bandas de CD próximas do limite de detecção dos espectrofotômetros de CD convencionais, em torno de 190 nm. Transições de alta energia são estudadas apenas em instrumentos à vácuo, mas sofrem fortes interferências dos solventes, fazendo com que tais estudos sejam limitados a filmes finos de sólidos.

Monossacarídeos têm sido extensivamente investigados, e algumas correlações conformacionais dos anéis podem ser extraídas em regiões do espectro de CD por volta de 170 nm. Mais uma vez, tais medições são limitadas, já que normalmente só podem ser feitas em CDs ligados a luz de síncrotron e também devido a interferência dos solventes.

O CD também tem sido bastante utilizado para estudo de carboidratos complexos como glicosaminoglicanos, heteropolissacarídeos compostos por um açúcar aminado (D-glicosamina ou D-galactosamina) unido por ligação glicosídica a um ácido urônico (D-glicurônico ou L-idurônico). Espectros de CD para diferentes glicosaminoglicanos podem ser observados na Figura 6-10.

As características de espectros de glicosaminoglicanos provêm predominantemente das transições eletrônicas $n \rightarrow \pi^*$ dos carboxilatos dos resíduos de ácido urônico e transições $\pi \rightarrow \pi^*$ dos cromóforos N-acetila dos resíduos de glicosamina. Em ambos os casos, a principal contribuição para as transições vem dos elétrons dos átomos de oxigênio. Para o ácido urônico, envolvem a função

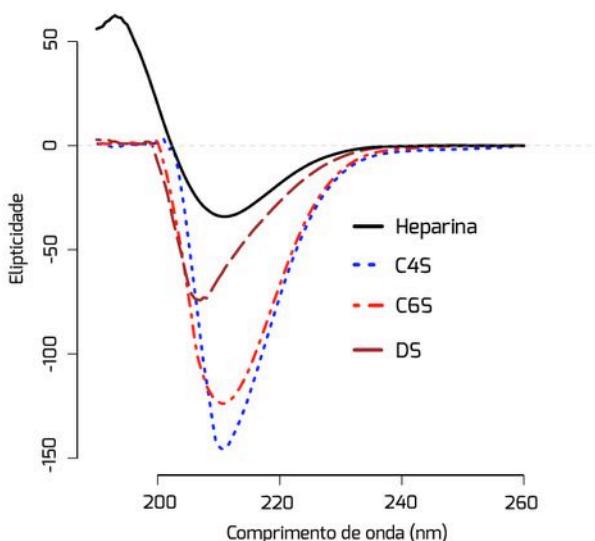



Figura 6-10: Espectro de CD de diferentes glicosaminoglicanos. C4S, condroitina 4-sulfatada; C6S, condroitina 6-sulfatada; DS, dermatam sulfato e heparina.

éter, a ligação glicosídica e as hidroxilas, produzindo uma banda positiva com valores máximos em torno de 190 nm. Para o grupo N-acetila e carboxilato, tem-se uma banda negativa com máximo em torno de 210 nm.

Como dito anteriormente, o CD pode ser utilizado para estudar a conformação de carboidratos e, no caso de glicosaminoglicanos, os resíduos de ácido urônico (β-D-glicurônico e α-L-idurônico) possuem bandas no espectro de CD de sinais opostos. Podem-se observar na Figura 6-10 os espectros de CD para DS, C4S e C6S, glicosaminoglicanos que contêm principalmente o ácido glicurônico.

Os espectros destes glicosaminoglicanos são peculiares, apresentando apenas uma larga banda negativa de máxima em torno de 210 nm. DS tem sua banda negativa ligeiramente deslocada à esquerda, com máxima em torno de 207 nm. Tal fenômeno pode ser explicado pelo fato de que ele também contém ácido idurônico. Além disso, a ausência da banda positiva de máxima em 190 nm pode refletir diferenças nas ligações glicosídicas já que DS, C4S e C6S apresentam β-D-galactosamina N-acetilada (ligação β), enquanto que a heparina contém α-D-glicosamina N-acetilada e/ou N-sulfatada (ligação α).

Ácidos nucleicos

As bases purínicas e pirimidínicas de DNA e RNA são, em grande parte, responsáveis pelo espectro de CD de ácidos nucleicos na faixa de comprimento de onda normalmente estuda por espectrofotômetros convencionais, uma vez que os carboidratos e grupos fosfato não absorvem significativamente acima de 200 e 180 nm, respectivamente.

Neste tipo de macromolécula, o CD é empregado principalmente no estudo da manutenção da geometria relativa das bases, pois cada uma possui um conjunto característico de transições $\pi\rightarrow\pi^*$ entre 180 e 300 nm.

Todas as cinco bases têm uma ou duas bandas de intensidade moderada, por volta de 260 nm, e várias bandas mais intensas, entre 180 e 200 nm. Além disso, cada base possui várias transições $n\rightarrow\pi^*$ entre 180 e 300 nm, porém de pequena absorbância. Embora potencialmente fortes no CD, as faixas de $n\rightarrow\pi^*$ não foram totalmente identificadas, sendo os espectros de CD de nucleosídeos, nucleotídeos e polinucleotídeos dominados pelas contribuições $\pi\rightarrow\pi^*$.

A estrutura $2^{ária}$ do DNA também pode ser estudada por CD (ver capítulo 2). A conformação B-DNA, encontrada normalmente em solução aquosa, tem uma banda positiva próximo 275 nm e uma banda negativa de magnitude similar perto de 245 nm (Figura 7A-10). Já a conformação A-DNA é favorecida pela adição de solventes orgânicos, geralmente etanol. No UV próximo, a transição B → A é marcada por um aumento significativo na banda positiva e diminuição na amplitude da banda de máxima em 245 nm. Outra característica é a presença de uma forte banda negativa em torno de nm 210 (Figura 7B-10). O C-DNA, por sua vez, apresenta banda intensa negativa por volta de 240 nm (Figura 7C-10).

Com base no espectro de CD, atribuições a um dos grupos de estrutura $2^{ária}$ po-

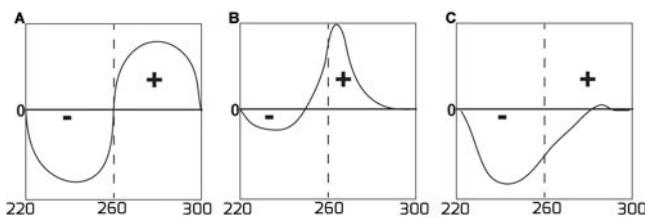



Figura 7-10: Representação esquemática dos espectros de CD para as diferentes estruturas secundárias de DNA.

dem ser feitas. Contudo, devido ao número considerável de subgrupos de estrutura 2ária e à dependência desta da sequência de nucleotídeos, informações detalhadas sobre a conformação do DNA não podem ser extraídas unicamente baseadas no espectro de CD.

Lipídeos

Aplicações de CD no estudo de lipídeos são raras, sendo sua mais frequente aplicação no estudo de proteínas de membrana em seu ambiente nativo, ou seja, inseridas na membrana. Porém, dois tipos de artefatos devem ser evitados. Suspensões de fragmentos de membrana podem induzir fortes efeitos de espalhamento de luz. Adicionalmente, eles podem apresentar espalhamento preferencial da luz circularmente polarizada à esquerda e à direita. Tal fenômeno se comporta como um sinal de CD, distorcendo o verdadeiro CD da proteína.

Ainda, fragmentos de membrana também distorcem os sinais de CD devido a um efeito conhecido como *Duysens' flattening*. Este efeito ocorre em amostras com uma distribuição não homogênea de cromóforos que estão associados com a fomação de micelas. Alguns métodos foram desenvolvidos buscando evitar tais dificuldades. Requerem, contudo, que a proteína de membrana seja transferida da sua membrana nativa para vesículas unilamelares que possuam, em média, apenas uma proteína por vesícula. Tais artefatos também podem ser evitados através da solubilização das proteínas em detergente não iônico, manobra esta que, contudo, pode induzir alterações conformacionais na proteína.

10.6. Situações práticas

Deconvolução espectral

A deconvolução espectral é utilizada para a resolução e/ou decomposição de um conjunto de sinais sobrepostos nos seus componentes separados através de algoritmos de ajuste de curva. Para a determinação da estrutura 2ária de proteínas, o espectro original é decomposto nos componentes hélice α, folhas β e estruturas irregulares e comparado a um banco de dados de proteínas com estrutura 2árias conhecidas.

No exemplo abaixo, o espectro de CD da albumina humana (Figura 8-10A) é decomposto nas suas estruturas 2árias componentes (Figura 8-10B) e, a partir destes, a proporção de cada tipo de estrutura calculada, totalizando 72% hélices α, 16% de folhas β e 12% de estruturas irregulares.

Interação proteína-ligante

Mudanças conformacionais sofridas por uma dada proteína após sua complexação a um determinado composto também podem ser determinadas por CD. Alterações na estrutura 2ária da proteína, promovidas por esta complexação, irão mudar o espectro de CD, de forma que algumas mudanças conformacionais podem ser detectadas.

É importante ressaltar que espectros de CD deverão ser coletados para todos os componentes do sistema em estudo, isto é, para a proteína e para o ligante em suas formas livres e para o complexo proteína-ligante. A partir destas medidas pode-se realizar subtrações espectrais, isto é, $CD_{proteína\text{-}ligante} - CD_{ligante}$. A partir destes dados é possível, por exemplo, comparar a capacidade de diferentes ligantes em modificarem o conteúdo de estrutura 2ária de uma determinada proteína receptora. Os espectros da proteína e da subtração serão deconvoluídos como descrito no item anterior.

No exemplo abaixo (Figura 9-10), pode-se observar o espectro da antitrombina humana livre e complexada a um composto

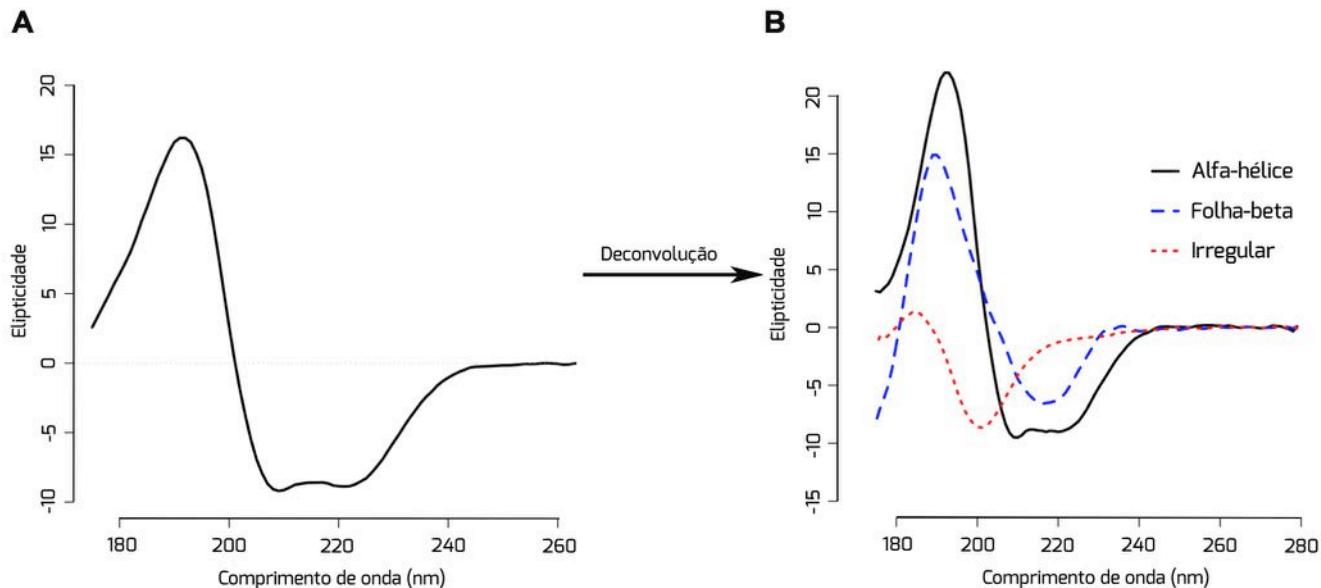



Figura 8-10: Deconvolução espectral esquemática da albumina sérica humana.

pentassacarídico, análogo da heparina de alta massa molecular empregada terapeuticamente. Após as devidas subtrações espectrais podemos determinar as mudanças induzidas pela ligação do pentassacarídeo à antitrombina, resultando em um aumento de 6,6% no conteúdo de hélices α e uma diminuição de 2% no conteúdo de folhas β e 2,5% no conteúdo de estruturas desordenadas.

CD e PCA

A análise de componentes principais (PCA, *Principal Component Analysis*) é um método matemático empregado para desvendar padrões em um conjunto complexo de dados (neste caso espectros de CD) e extrair informações cruciais, eliminando assim possíveis fontes de ruído.

A combinação linear que extrai a variância máxima dos dados é denominada de componente principal. Uma vez que ela é encontrada, é removida e o processo repetido para identificar o próximo componente principal. Isso se repete até que toda a variância dos dados seja explicada, fato que na prática não ocorre devido ao ruído residual.

Na análise de PCA, os componentes representam as dimensões subjacentes que resumem ou explicam um conjunto original de dados observados. *Component loadings* são os coeficientes de correlação entre as variáveis e os fatores. Os *components loadings* ao quadrado indicam a percentagem de variância da variável original. *Component scores* representam uma medida composta criada para cada observação em cada fator extraído da análise fatorial.

A Figura 10-10 mostra que a análise matemática dos espectros de CD é eficaz na diferenciação de glicosaminoglicanos, heparina e seus derivados. As características estruturais que são introduzidas nas heparinas de

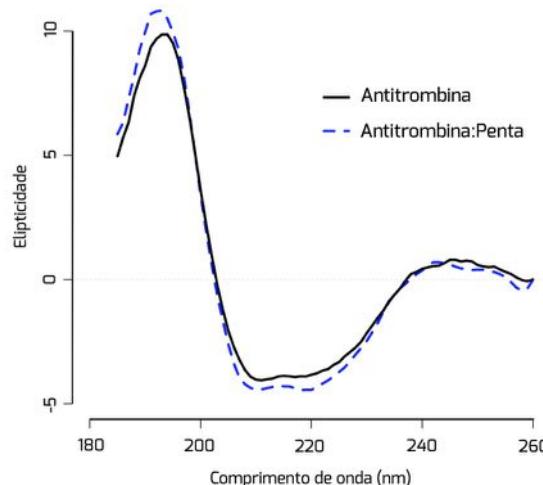



Figura 9-10: Espectro de CD da antitrombina humana (linha preta) e do complexo antitrombina:pentassacarídeo (linha azul).

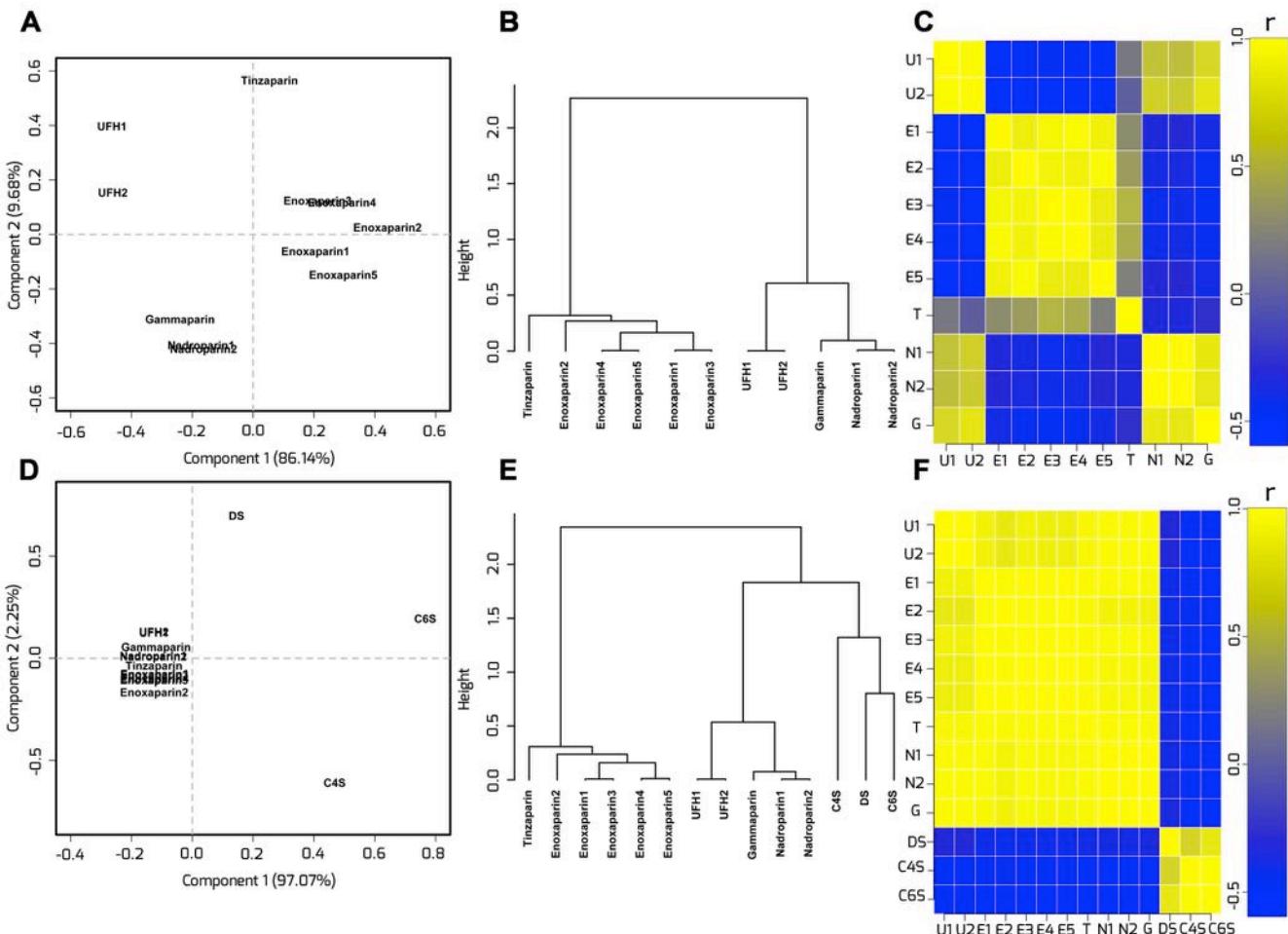



Figura 10-10: Análise matemática dos espectros de CD de glicosaminoglicanos. (a e d) *Loading plot*. (b e e) Análise de cluster. (c e f) Matriz de correlação. U, heparina não-fracionada; E, enoxaparina; T, tinzaparina, N, nadroparina, G, gammaparina; DS, dermatam sulfato, C4S, condroitina 4-sulfatada; C6S, condroitina 6-sulfatada; r, coeficiente de correlação. Imagem extraída com permissão de Lima e colaboradores, *Low molecular weight heparins: Structural differentiation by spectroscopic and multivariate approaches*, *Carbohydr. Polymers*, **2011**, *85*, 903-909, 10.1016/j.carbpol.2011.04.021.

baixo peso molecular ao longo das reações de despolimerização química e enzimática, bem como diferenças nos tipos de ligação glicosídica, N-acetilação, padrão de N- e O-sulfatação e composição monossacarídica resultam em características especificas nos seus espectros de CD que são facilmente diferenciadas pela análise matemática dos dados.

Aquisição de um espectro de CD

i) Evitar tampões quirais e que possuem forte absorção no UV, principalmente na faixa entre 180-260 nm;
ii) Filtrar todas as soluções, inclusive a amostra a ser estudada, evitando assim a presença de partículas causadoras de espalhamento de luz;
iii) Antes de coletar o espectro para a amostra em estudo é importante coletar um branco que nada mais é que o espectro do tampão;
iv) Em experimentos comparativos, usar sempre as mesmas condições experimentais, tais como temperatura, tampão utilizado, concentração dos componentes, comprimento do caminho óptico e resolução (ou seja, frequência de intervalos, em nm, na qual é feita a aquisição dos dados);
v) Para proteínas, é importante coletar espectros em diferentes concentrações

e observar se há mudança nos sinais. Havendo mudanças, a proteína em estudo está agregando;

vi) Para açúcares, é importante mantê-los na mesma forma catiônica, uma vez que diferentes contra-íons produzirão espectros distintos.

10.7. Conceitos-chave

Análise de componentes principais: ferramenta matemática que desvenda padrões em um conjunto de dados complexos.

Coeficiente de absorbância: capacidade de um mol de uma dada substância em absorver luz em um determinado comprimento de onda.

Dicroismo circular: é a medida da absorbância diferencial entre as duas rotações de luz circularmente polarizada por uma molécula assimétrica.

Duysens' flattening: distribuição não homogênea de cromóforos em uma dada molécula.

Enantiômeros: imagens especulares (isto é, geradas a partir da reflexão em um espelho), não sobreponíveis, de uma determinada molécula, que assim apresenta a propriedade de quiralidade.

Lei de Lambert-Beer: é uma relação, determinada empiricamente, entre a luz absorvida por um determinado material e propriedades intrínsecas a este material.

Quiralidade: propriedade de uma molécula não ser sobreponível a sua imagem especular.

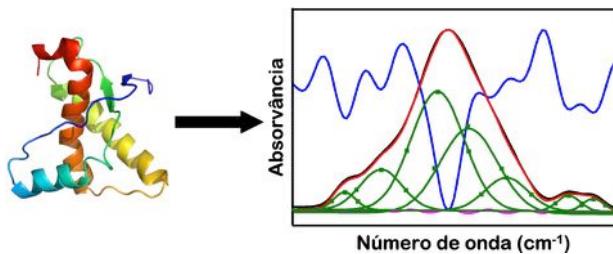
Vesículas unilamelares: Formas lipossomais constituídas por apenas uma bicamada fosfolipídica.

10.8. Leitura recomendada

PURDIE, Neil; BRITTAIN, Harry G (Org.). ***Analytical Applications of Circular Dichroism***. Amsterdam: Elsevier Science Limited, 1994.

FASMAN, Gerald D. (Org.) ***Circular Dichroism and the Conformational Analysis of Biomolecules***. New York: Plenum Press, 1996.

WALLACE, B. A. Conformational changes by synchrotron radiation circular dichroism spectroscopy. ***Nat. Struct. Biol.*** 7, 708–709, 2000.

RODGERS, David S. ***Circular Dichroism: Theory and Spectroscopy***. Hauppauge: Nova Science Publishers, 2011.

11. Infravermelho

Estrutura 3D da proteína prion de camundongo e seu espectro de infravermelho na região da amida I.

Yraima Cordeiro
Luís Maurício T. R. Lima

11.1. Introdução

O espectro eletromagnético é composto por diferentes tipos de radiações, dos raios gama (maior energia) às ondas de rádio (menor energia, Figura 1-11). Entre estes extremos de radiações, diversos tipos de ondas possuem aplicações ao estudo de biomoléculas, como os raios-X (ver capítulo 13), o ultravioleta (ver capítulo 10) e o infravermelho, assunto deste capítulo.

A região do infravermelho (IV) no espectro eletromagnético (Figura 1-11) está compreendida entre aproximadamente 14.000 cm^{-1} e 200 cm^{-1}, indo do que chamamos IV próximo ao IV distante, respectivamente. Adicionalmente, a região compreendida entre 4.000 e 400 cm^{-1} (2.500 a 25.000 nm) é denominada IV médio, e possui destaque nos estudos da estrutura 2ária de proteínas.

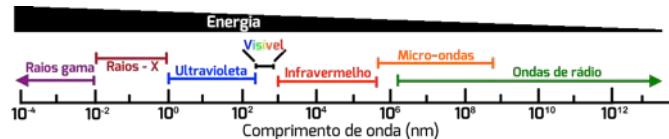



Figura 1-11: Esquema das diferentes regiões do espectro eletromagnético. Quanto maior o comprimento de onda, menor a energia da radiação.

Medidas empregando IV vêm sendo aplicadas há décadas na análise e caracterização de pequenos compostos orgânicos e, para tal, existem diversos livros texto disponíveis. Este capítulo se dedica, contudo, a aplicações mais recentes, focadas no estudo de biomacromoléculas. Mesmo que o princípio da técnica seja o mesmo, as diferenças em ordens de grandeza no número de átomos envolvidos trazem à tona uma série de particularidades, que veremos em seguida.

Quando incidimos uma determinada radiação sobre a amostra em estudo, as moléculas ali contidas absorvem energia. Esta energia promove a passagem dos elétrons de um estado fundamental (E_0) a um estado de maior energia (E_1). Após o desligamento da fonte de luz, os elétrons retornam a E_0 depois de alguns segundos, liberando a energia absorvida. Esta energia, por exemplo, pode estar na região do ultravioleta permitindo, por exemplo, medições de dicroismo circular (ver capítulo 10) e de fluorescência.

Entretanto, a absorção de energia radi-

ante não envolve somente transições eletrônicas, mas a energia total da molécula (E_{total}). Esta energia pode ser representada pelo somatório das energias associadas a: 1) rotação da molécula na solução (E_R), 2) movimento dos átomos dentro da molécula, constituindo a energia vibracional (E_V), e 3) movimento dos elétrons ao redor do núcleo, a chamada energia eletrônica (E_E). Assim, podemos representar $E_{total} = E_R + E_V + E_E$. Dependendo do nível de energia da radiação incidente, quando a molécula retorna de seu estado excitado para o estado fundamental, também há perda nas energias de vibração (E_V) e rotação (E_R).

Assim, nos comprimentos de onda abaixo de 25 μm (400 cm^{-1}), ou seja, em torno da região do IV médio, a radiação tem energia suficiente para provocar modificações nos níveis de energia vibracional (E_V) da molécula, e estas modificações são acompanhadas por alterações nos níveis de energia rotacional (E_R). Isto ocorre quando a luz no IV coincide com a energia necessária para que ocorra uma determinada vibração molecular.

Ao estudar as mudanças no comportamento molecular após a incidência de radiação IV, podemos caracterizar os diferentes modos de vibração e rotação de uma molécula, os quais constituem o espectro de infravermelho.

Análises na região do IV permitem descrever o arranjo espacial dos átomos nas moléculas do composto em estudo, ou seja, como é a sua estrutura química; fornecem informações sobre comprimento e a força de ligações químicas; fornecem evidências para o comportamento químico ou físico relativo de uma molécula (estado redox, catálise enzimática e fosforilação, dentre outras), além de permitirem a análise qualitativa e quantitativa de uma determinada molécula.

Para compreendermos como o espectro de IV pode fornecer informações sobre o arranjo molecular de um determinado composto e sobre a interação deste com o ambiente, devemos definir a frequência de vibração de um oscilador diatômico. Esta frequência (ν) pode ser representada por:

$$\nu = (k/m_r)^{0,5}/2\pi$$

onde k é a constante de força entre os dois átomos e m_r a massa reduzida.

De forma simplificada, a massa reduzida (m_r) é um termo utilizado em mecânica Newtoniana ao se estudar um sistema diatômico (ou seja, no qual há interação entre dois átomos). A m_r engloba a massa do primeiro e do segundo átomos, simplificando um sistema de dois componentes em um sistema de um componente.

Esta equação nos diz que a frequência de vibração aumenta quanto maior for a força de interação entre os dois átomos (isto é, a força da ligação química). Em outras palavras, quando aumenta a densidade eletrônica na ligação entre os dois átomos (de uma ligação simples para uma ligação dupla e para uma ligação tripla) aumenta a frequência de vibração. Dessa forma, qualquer fator inter- ou intramolecular que altere a densidade eletrônica nas ligações (como o tipo de átomo) irá afetar o espectro vibracional obtido por IV. E quanto maior for a massa dos átomos, mais lenta será a vibração (menor frequência).

Se pensarmos em ligações O-H e N-H, embora sejam ambas ligações simples, o átomo de oxigênio é mais eletronegativo que o átomo de nitrogênio. Assim, a ligação O-H é mais polar que a ligação N-H, resultando em uma força de interação diferente entre os átomos e, por conseguinte, uma vibração diferente. Adicionalmente, como veremos adiante, o espectro de IV não é definido somente por características intramoleculares do composto em estudo, mas também de interações com outras moléculas.

Com a absorção da luz no IV as ligações atômicas vibram, promovendo deformações axiais (estiramentos) ou angulares (dobras). Estiramentos são alongamentos da ligação química, enquanto deformações angulares são dobras nesta ligação química. Os estiramentos e deformações podem ser simétricos ou assimétricos, como representado na Figura 2-11 para a molécula de água. As deformações angulares simétricas que ocorrem no plano são chamadas de deformação em tesoura, enquanto que as deformações assi-

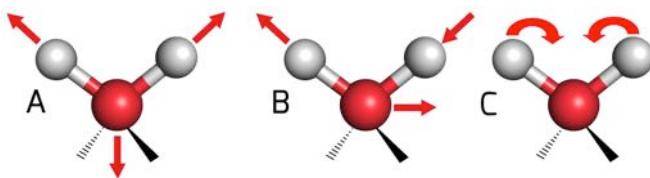



Figura 2-11: Modos vibracionais da H_2O. As setas vermelhas indicam em A, estiramento simétrico; B, estiramento assimétrico; C, deformação angular no plano (em tesoura).

métricas no plano são chamadas de vibrações em balanço ou rotação.

Existem também deformações que ocorrem fora do plano, que podem ser denominadas como deformações em balanço (simétrico) ou em torção (assimétrico, saindo ou entrando da tela deste computador, por exemplo). Na literatura, muitas vezes estas deformações são representadas como ν (deformação axial) e σ (deformação angular).

Como representado na Figura 3-11, é possível notar que espectros de IV podem ser extremamente complexos, visto a quantidade de estiramentos e deformações angulares que podem estar presentes em uma molécula relativamente pequena. Tomemos como um exemplo a molécula de ureia que, embora tenha somente três ligações químicas diferentes (isto é, C=O, N-H e C-N), apresenta mais de 7 picos em seu espectro IV (Figura 3-11).

Além do número de picos (ou bandas) em um espectro de IV, a intensidade de cada banda varia de acordo com a quantidade de luz absorvida por determinada ligação na frequência observada. Dessa forma, há picos ou bandas fracos (baixa intensidade) e picos ou bandas fortes (alta intensidade) em espectros de IV da maioria das moléculas (ver picos no espectro IV da molécula de ureia, Figura 3-11).

Como podemos verificar na Figura 4-11, as frequências vibracionais de ligações químicas presentes em proteínas estão presentes em diversas regiões do espectro de IV. Para pequenos compostos, a análise dos espectros de IV pode fornecer informações sobre o arranjo espacial dos átomos envolvidos.

Entretanto, para macromoléculas, que são o foco deste capítulo, há obviamente uma

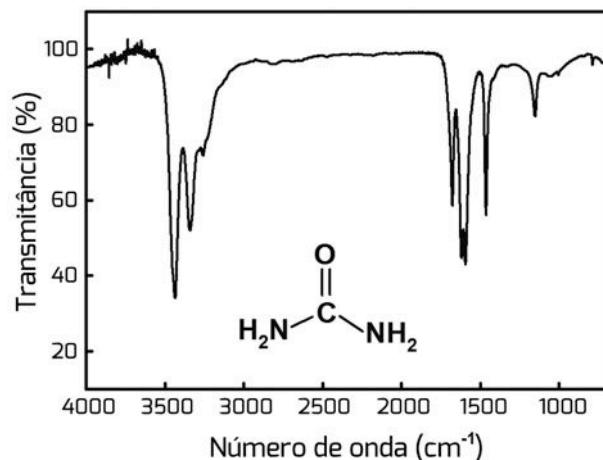



Figura 3-11: Espectro de infravermelho da ureia.

grande sobreposição de frequências vibracionais. Dessa forma, não é possível determinar a estrutura molecular de uma proteína por IV. Podemos, contudo, obter informações sobre seus componentes de estrutura $2^{ária}$ e seu grau de enovelamento.

A análise de estrutura $2^{ária}$ de proteínas e de outras macromoléculas biológicas por infravermelho teve início na década de 1970. Com o advento de espectrofotômetros de IV não-dispersivos (FTIR) e novos detectores, houve uma melhoria significativa na qualidade e conteúdo de informação a ser obtido de espectros de infravermelho de proteínas.

11.2. Instrumentação

A notação mais utilizada para análise no IV é dada em números de onda. Esta notação é uma grandeza física diretamente proporcional à energia da radiação eletromagnética e, portanto, inversamente proporcional ao comprimento de onda em nanômetros. A unidade da notação em números de onda é centímetros recíprocos ou cm^{-1}.

O número de onda pode ser definido como o número de ondas da radiação eletromagnética que são comportados dentro de um espaço de 1 cm (Figura 5-11). Por exemplo, uma radiação com comprimento de onda de 300 nm equivale a 33,333 cm^{-1}, e uma radiação com comprimento de onda de 500 nm

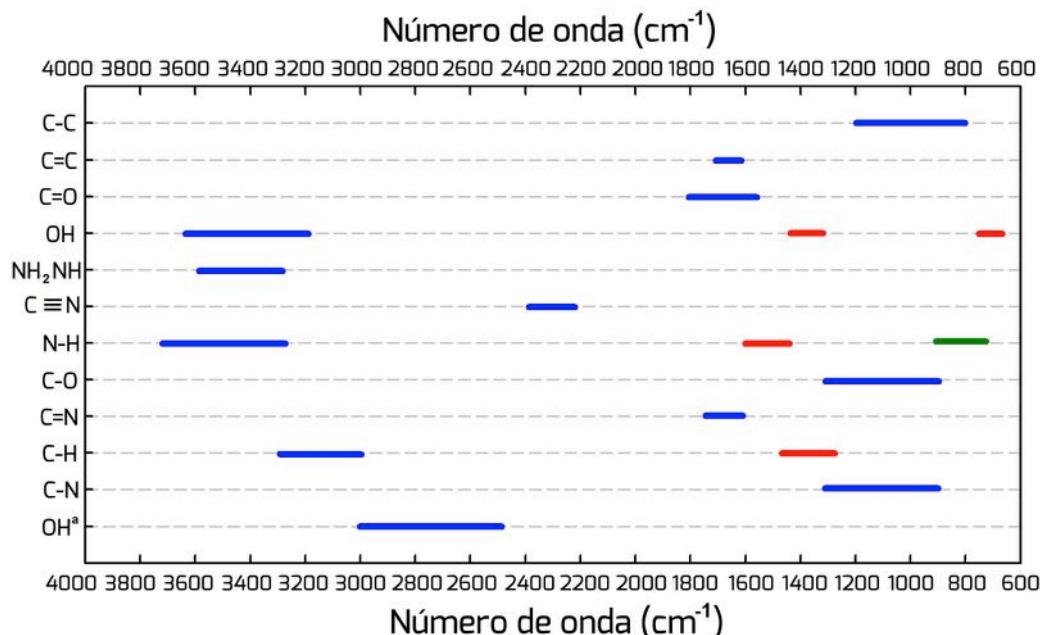



Figura 4-11: Frequências de absorção no IV de algumas ligações químicas. Estão representadas frequências vibracionais resultantes de estiramentos (azul), dobras ou deformações em tesoura (vermelho) e em balanço (verde) da ligação.

(menos energética do que a primeira) possui um comprimento de onda de 2.000 nm. Assim, como o número de onda é diretamente proporcional à energia e, portanto, à frequência, quanto maior o valor em números de onda, mais alta será a frequência daquela radiação eletromagnética.

A energia da radiação eletromagnética é definida por:

$$E = h\nu = hc/\lambda$$

onde h é a constante de Planck (6,6261 x 10^{-34} J), c é a velocidade da luz no vácuo (2,99792 x 10^{8} m/s), ν é a frequência da radiação (dada por $\nu = hc/\lambda$) e λ é o comprimento de onda em nanômetros.

Para conversão da notação de frequências de absorção no IV entre nanômetros e números de onda, considerando-se que 1 cm = 10.000.000 nm (10^{7}), então:

$$\text{número de onda} = 1/\lambda.10^{7}$$

Antes de discutirmos sobre a análise de espectros de IV de proteínas, faremos uma breve explicação sobre a instrumentação empregada nestes estudos. O equipamento básico consiste em uma fonte geradora de luz no IV, de espelhos organizados para direcionar a luz para a amostra e de um detector para captar a luz transmitida. A fonte geradora de IV é, em geral, composta por óxidos de terras raras (por exemplo, carbeto de silício), que emitem radiações na região do IV quando aquecidos a altas temperaturas (1.000 a 1.800 ºC).

Espectrômetros de IV por transformada de Fourier contém um dispositivo chamado de interferômetro. O interferômetro é um sistema óptico capaz de fornecer uma radiação aproximadamente monocromática na região de 2,5 μm a 15 μm ou até 50 μm. O interferômetro permite a separação e depois a recombinação do feixe de infravermelho, a partir da passagem da luz pelo separador do feixe (*beam splitter*) e a incidência de cada

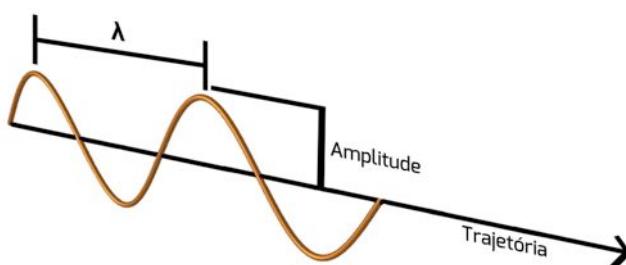



Figura 5-11: Representação esquemática de uma onda eletromagnética.

feixe resultante sobre um espelho fixo e um espelho móvel. O sinal de saída é chamado de interferograma (Figura 6-11).

O funcionamento de um interferômetro consiste na passagem do feixe luminoso pelo separador de feixe (B), e parte do feixe é refletido pelo espelho móvel (EM) e retorna ao separador. O outro feixe é refletido do separador e, então, pelo espelho fixo (EF), retorna a B. O feixe recombinado sai do interferômetro, passa através da amostra (A) e viaja até o detector (D) (Figura 6-11). O sinal é captado a intervalos precisos, correspondentes a passos iguais na diferença de caminho óptico (ou seja, a distância da trajetória da luz pela amostra), resultando em um sinal combinado de interferência destrutiva e construtiva em função das diferenças de fases (ver abaixo), o que origina o nome do dispositivo e do sinal obtido. O interferograma é resultante do registro do sinal no detector em função da diferença de caminho entre os dois feixes. Como referência, é utilizado um laser de hélio-neônio, e sua radiação monocromática de 632,8 nm atravessa o mesmo caminho óptico do feixe de IV.

A varredura em FTIR corresponde ao deslocamento mecânico do espelho móvel (E_M). Quando a distância B - E_M é igual à distância B - E_F, os dois feixes refletidos percorrem a mesma distância, estando totalmente em fase (ver adiante). Como resultado, os dois feixes interferem construtivamente, e o detector observa um máximo de intensidade. Esta posição do espelho móvel é chamada de diferença zero de caminho óptico (*zero path difference* ou ZPD). Neste caso 2.(B - E_M) = 2.(B - E_F). À medida que E_M afasta-se do ZPD, a distância B - E_M aumenta em relação à distância B - E_F. Quando os dois feixes estiverem 180º fora de fase, e a interferência será destrutiva, provocando um mínimo na resposta do detector.

Figura 6-11: Esquema de um interferômetro. A luz no IV, gerada pela fonte, trafega até o separador do feixe (B), que é separado e incide sobre o espelho fixo (E_F) e sobre o espelho móvel (E_M). O feixe é recombinado em B, atravessa a amostra (A) e chega ao detector (D).

O espectro resultante (dados no domínio de frequência) é a solução de Fourier para o sinal do interferograma (dados no domínio de tempo). Espectrômetros FTIR permitem medidas mais rápidas do que os antigos espectrômetros, denominados dispersivos (Tabela 1-11).

Para entendermos o significado de diferença de fase vamos tomar como exemplo duas radiações (isto é, ondas eletromagnéticas) que apresentam a mesma frequência e, portanto, a mesma energia. Se ambas estão trafegando ao mesmo tempo no espaço, estas ondas estão em fase e há um somatório de suas amplitudes (ver Figura 7-11).

Se há um retardo de uma das frequências em relação à outra, estas ondas estão agora fora de fase. Se as ondas estão 180º fora de fase a interferência é destrutiva, pois o somatório das ondas resulta em 0. Em contrapartida, se estão em fase a interferência é construtiva. Esta mesma definição pode ser aplicada para a vibração das ligações químicas presentes em uma dada molécula, as quais podem estar vibrando em fase ou fora de fase

11.3. Vibrações de H_2O e ^2H_2O

Água no estado líquido e vapor de água interferem de forma intensa em espectros de IV de proteínas. As principais frequências vibracionais da água (Tabela 2-11) se sobrepõem à região da amida I, principal banda no IV que dá informações sobre a estrutura 2ª^ria^ de proteínas.

Sendo assim, para se realizar medidas de proteínas em solução, as amostras são

Tabela 1-11: Diferenças entre espectrômetros por transformada de Fourier (FTIR) e espectrômetros dispersivos.

IR dispersivo	FTIR
Partes móveis: desgaste e tolerância mecânica	Somente 1 espelho se movimenta durante coleta
Pequena fração de v é detectada por unidade de tempo. Varredura completa em 10 – 15 min	Todos os valores de v são detectadas simultaneamente. Espectro coletado < 1 s
Baixa velocidade de varredura	Rápida velocidade de varredura: cinética
Não há referência interna para verificar a exatidão de v, exigindo calibração com espectros referência	Uso de He-Ne: sistema de calibração interno com exatidão e precisão na faixa de 0,01 cm^{-1}
Amostra localizada próximo à fonte, gerando possíveis problemas térmicos	Amostra localizada longe da fonte

usualmente diluídas em D_2O (^2H_2O), ou óxido de deutério. Como o deutério apresenta massa maior do que o hidrogênio, sua frequência vibracional é menor do que a da H_2O, não havendo mais sobreposição na região da amida I, onde são vistas hélices α e estruturas desordenadas (ver a seguir). Assim, quando temos ^2H ao invés de ^1H, as principais bandas vibracionais da água líquida são deslocadas para frequências mais baixas.

O espectro de IV da água no estado líquido sofre alterações dependentes das ligações de hidrogênio o que, por sua vez, não ocorre no espectro IV da água em vapor (onde estas interações estão ausentes). Para a água no estado líquido, com o aumento da força das ligações de hidrogênio observa-se o deslocamento das deformações axiais e das deformações angulares para menores e maiores frequências, respectivamente. Estas variações na intensidade das ligações de hidrogênio podem ocorrer, por exemplo, devido a mudanças na temperatura. Neste caso, um aumento na temperatura enfraquece as ligações de hidrogênio, fortalecendo a ligação

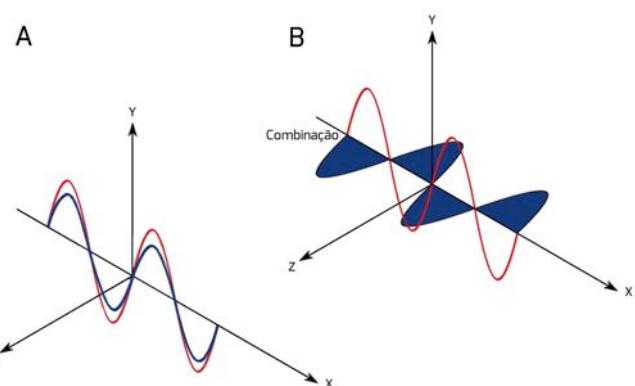



Figura 7-11: Exemplo esquemático de duas ondas em fase (A) e duas ondas 180º fora de fase (B).

covalente O-H que passa a vibrar em frequências maiores.

11.4. Realizando medidas de IV

Como vimos acima, há uma grande sobreposição entre vibrações da molécula de água com a região do espectro de IV empregada no assinalamento das estruturas $2^{árias}$ de proteínas. Assim, precisamos reduzir ao máximo o conteúdo de H_2O da amostra a ser analisada.

Para medidas em solução, uma alternativa é realizar todas as etapas de obtenção da proteína de interesse em ^2H_2O. Entretanto, esta alternativa não é usualmente viável devido ao alto custo da ^2H_2O e, ainda, por este se hidratar rapidamente.

Uma abordagem alternativa e amplamente utilizada é obter a proteína normalmente (estratégia de purificação normal, em solvente aquoso), remover toda a H_2O por secagem (sublimação da água por liofilização ou outra técnica de escolha), ressuspender o material seco em ^2H_2O, secar a amostra novamente para permitir a troca de ^1H por ^2H e ressuspender a amostra em ^2H_2O em uma concentração maior que 1% massa/volume para a realização da medida. A amostra em solução é aplicada entre duas janelas (duas "fatias") formadas por material transparente ao IV médio, como fluoreto de cálcio (CaF_2), por exemplo, que são montadas em um porta-amostras (Figura 8-11).

Caso não se deseje realizar medidas em

Tabela 2-11: Principais vibrações de $^{1}H_2O$ e $^{2}H_2O$ (D_2O) na região do IV.

Vibração	H_2O líquida (25 °C)		D_2O líquido (25 °C)	
	ν (cm^{-1})[a]	E_0 ($M^{-1}.cm^{-1}$)[b]	ν (cm^{-1})[a]	E_0 ($M^{-1}.cm^{-1}$)[b]
Dobra	1.643,5	21,8	1.209,4	17,4
Combinação de dobra e oscilação	2.127,5	3,50	1.555,0	1,91
Estiramentos simétricos e assimétricos	3.404,0	99,9	2.504,0	71,5

[a] ν, frequência vibracional; [b] E_0, coeficiente de extinção molar.

solução, é possível analisar a amostra seca na forma de pastilha com brometo de potássio (KBr). KBr é transparente na região do infravermelho médio, e é também o componente do separador do feixe no interferômetro. Em linhas gerais, mistura-se a amostra de interesse a 1% com KBr (1 mg da amostra para 100 mg de KBr, por exemplo) em um gral com um pistilo de quartzo e, por pressão mecânica, gera-se um disco da amostra com espessura de ~10 mm que é acondicionado ao porta-amostras do equipamento para realização da leitura. É importante realizar uma maceração eficiente da amostra com KBr, para resultar em uma distribuição uniforme da sua amostra com o pó.

Para a amostra seca, é ainda possível realizar medidas empregando técnica de reflectância total atenuada (*attenuated total reflectance*, ATR). Nesta técnica, a amostra sólida é depositada sobre um cristal de índice de refração maior que a amostra e comprimida sobre esta superfície, de modo a impedir a presença de ar e água que poderiam atrapalhar a medida. A luz IV é então refletida sobre esta superfície. O feixe emerge do cristal (neste caso, é chamado de onda evanescente) e incide sobre a amostra, havendo absorção, refletindo de volta e sendo por fim redirecionada ao detector. Existe grande popularidade neste método devido à vantagem de não demandar pastilhamento e requerer apenas alguns microgramas de amostra seca.

Após o preparo da amostra, coleta-se inicialmente um espectro base (*background*) na ausência de amostra. Este espectro base normalmente é chamado de espectro de feixe único (*single-beam*), pois reflete a resposta em todas as frequências da região do IV médio (que é gerada pela maioria dos equipamentos de IV) sem nenhuma correção. Um espectro de feixe único de uma amostra pode ser corrigido pelo espectro base, o que irá gerar o espectro final de IV.

Contudo, medidas envolvendo proteínas requerem instrumentação com sensibilidade maior do que aquela empregada para pequenas moléculas, visto que o sinal da amida é mais fraco (baixa intensidade) devido à baixa absorção de luz no IV médio.

Antes de iniciarmos a coleta de um espectro de IV, devemos resfriar o detector com nitrogênio líquido (-196 ºC). Detectores MCT (mercúrio, cádmio e telureto) apresentam alta sensibilidade e são a escolha para análise de proteínas. Estes detectores semicondutores de fótons no IV são refrigerados para reduzir o ruído e o vazamento de corrente resultante dos processos de geração térmica. Detectores MCT operam a temperaturas de 80 a 200 K.

Mesmo para amostras medidas no es-

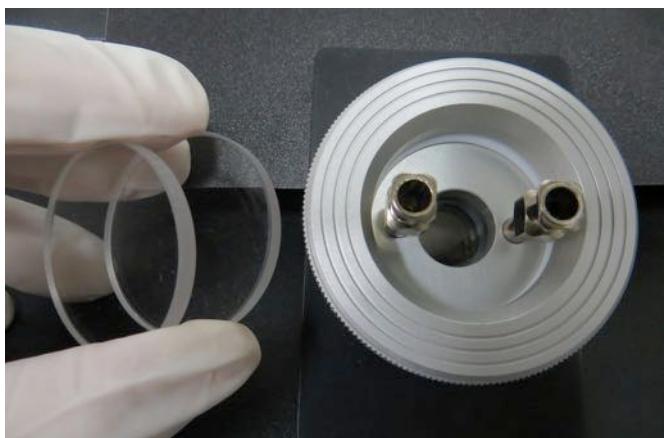



Figura 8-11: Janelas de fluoreto de cálcio (esquerda) e porta-amostra (direita). Dimensões típicas das janelas de CaF_2: 32 mm de diâmetro e 3 mm de espessura.

tado sólido (sem água líquida), deve-se efetuar a purga da região do porta-amostras com N_2 ou ar seco, pois vapor de água também absorve na região do IV médio e pode comprometer a análise da banda amida I (ver adiante).

Para realizar medidas de espectroscopia de IV por transformada de Fourier (FTIR) o ideal é coletar o maior número de varreduras possíveis, com resolução alta (de 1 a 2 cm^{-1}). O espectro resultante pode ser na escala de transmitância ou absorbância (Figura 9-11). Caso o espectro contenha muito ruído, é aconselhável diminuir a resolução da medida (por exemplo, 4 cm^{-1}) e/ou aumentar a quantidade de amostra analisada (aumentar a massa, caso depositada em cristal de ATR, ou aumentar a concentração, caso esteja medindo proteína em solução).

11.5. Espectros de IV de proteínas

A análise de estrutura $2^{ária}$ de proteínas a partir de seu espectro vibracional vem sendo realizada desde o início da década de 1980. É possível inferir se a proteína adota uma estrutura rica em hélices α, folhas β, ou se não apresenta estrutura $2^{ária}$ definida (ver capítulo 2), a partir da análise da banda amídica I de proteínas na região do IV médio. Além da amida I, o espectro vibracional de proteínas apresenta outros componentes que serão apresentados a seguir.

Como já descrito no capítulo 2, o estabelecimento de redes de ligação de hidrogênio entre resíduos de aminoácidos é um dos fatores que distingue os tipos de estrutura $2^{ária}$ adotadas por sequências polipeptídicas. Cada tipo de estrutura $2^{ária}$, por sua vez, implicará na adoção de valores para os ângulos φ e ψ ao redor da ligação peptídica. Estas interações afetam a frequência vibracional de ligações ente átomos, e isso será refletido no espectro de IV da proteína estudada. Dessa forma, é possível inferir que tipo de estrutura $2^{ária}$ a proteína analisada apresenta.

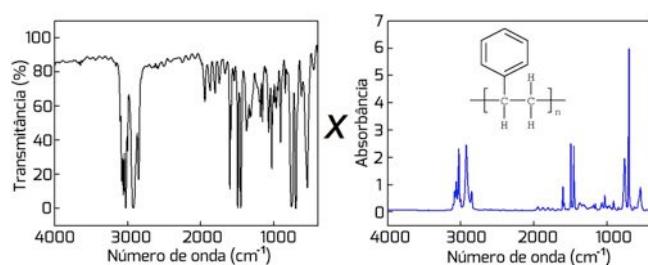



Figura 9-11: Absorção de poliestireno (estrutura no gráfico à direita) em filme na região do infravermelho médio. Na esquerda está o espectro em unidades de transmitância e, na direita, o mesmo espectro em unidades de absorbância.

Regiões vibracionais de proteínas

Parte do estudo das vibrações no IV da ligação peptídica (ou ligação amídica) foi baseado na análise dos componentes vibracionais da N-metil acetamida (NMA, Figura 10-11). Esta molécula é utilizada como composto modelo para definição de componentes vibracionais em proteínas, já que é a menor estrutura que contém um grupamento peptídico em E (ligações peptídicas em trans, as quais ocorrem na quase totalidade das proteínas).

As diferentes regiões vibracionais de proteínas no espectro de IV são chamadas de bandas amídicas ou amidas, pois resultam das diferentes interações realizadas pelos átomos que compõem a ligação amídica (ligação peptídica) com moléculas do solvente e com átomos da própria proteína, sejam estes da cadeia lateral ou do esqueleto polipeptídico (Figura 11-11). Por exemplo, como vimos no capítulo 2, a estrutura $2^{ária}$ de proteínas é mantida principalmente por ligações de hidrogênio entre os grupamentos N-H e C=O da cadeia polipeptídica com os mesmos grupamentos na volta seguinte da hélice ou na fita vizinha da folha.

Além das vibrações da cadeia polipeptídica (que informam sobre a estrutura $2^{ária}$ da proteína), vibrações das cadeias laterais de resíduos de aminoácidos também contribuem para o espectro de IV de proteínas. Entretanto, há uma grande sobreposição das vibrações de cadeias laterais, e algumas absorvem

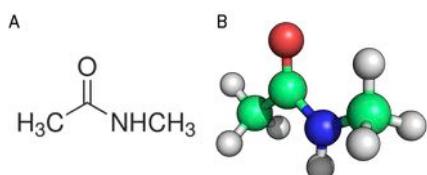



Figura 10-11: Estrutura da N-metil acetamida (NMA).

fracamente a luz IV. Portanto, é difícil identificá-las isoladamente. A seguir serão apresentadas as principais regiões vibracionais de proteínas e quais informações podem ser obtidas de cada uma destas regiões.

Amidas A e B

Estas bandas são resultantes do estiramento da ligação N-H e estão presentes na faixa de ~3.300 e ~3.170 cm^{-1}. Esta região é insensível à conformação da cadeia polipeptídica, e sua frequência depende da força da ligação de hidrogênio realizada pelo grupamento.

Amida I

Esta é a principal banda vibracional de proteínas, pois fornece informações sobre a estrutura $2^{ária}$ destas macromoléculas. A frequência média da amida I ocorre em torno de 1.650 cm^{-1}, e resulta principalmente do estiramento simétrico da carbonila ($\nu_{C=O}$), com pequenas contribuições da vibração C-N fora de fase, da deformação C-C-N e da torção N-H no plano. A estrutura do esqueleto polipeptídico irá determinar como as várias coordenadas internas irão contribuir para a vibração desta banda. A despeito de ser influenciada pela estrutura $2^{ária}$, esta vibração é muito pouco afetada pela natureza das cadeias laterais.

Amida II

A absorção da banda amida II ocorre em ~1.550 cm^{-1} quando o solvente utilizado no experimento de IV é H_2O. Esta vibração é a combinação fora de fase da torção N-H no

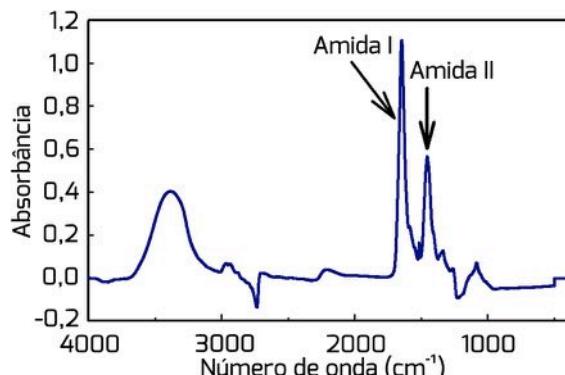



Figura 11-11: Espectro de absorção no IV médio de uma amostra proteica. Observe as regiões de amida I (1.700 a 1.600 cm^{-1}) e amida II (1.600 a 1.450 cm^{-1}).

plano e do estiramento da ligação C-N, com poucas contribuições da torção C-O no plano e das vibrações de C-C e N-C. Como para a amida I, esta vibração é pouco afetada pelas vibrações das cadeias laterais, mas a correlação entre estrutura $2^{ária}$ e frequência, nesse caso, é menos direta do que para a vibração amídica I.

Entretanto, a análise desta banda vibracional fornece informações a respeito do enovelamento proteico e sua dinâmica conformacional em experimentos de troca de ^1H por ^2H (troca hidrogênio – deutério), pois há um desvio da amida II para 1.450 cm^{-1} quando a proteína é diluída em ^2H_2O. Sendo assim, é possível acompanhar a troca de hidrogênios lábeis (como hidrogênios da ligação N-H da cadeia polipeptídica) por deutério durante tratamento térmico da proteína, interação com algum ligante e aumento na pressão, dentre outras variáveis. Átomos de hidrogênio em regiões mais protegidas da proteína irão demorar mais para trocar por deutério do que átomos de hidrogênio em regiões expostas

Há ainda uma terceira banda relacionada à ligação peptídica, a chamada banda de amida III. Esta banda, no NMA, é a combinação em

fase da dobra da ligação N-H e do estiramento da ligação C-N, principalmente. Em polipeptídeos, a composição dessa banda é mais complexa, pois depende da estrutura das cadeias laterais e a dobra do N-H contribui para várias bandas na região de 1.400 a 1.200 cm^{-1}. Como essas contribuições variam bastante, esta vibração é de pouca utilidade para análise de estrutura $2^{ária}$.

Vibração do esqueleto peptídico

Esta vibração ocorre de 1.200 a 880 cm^{-1} e resulta do estiramento das três ligações do esqueleto polipeptídico. Para o composto modelo NMA, estas vibrações geram duas bandas bastante definidas, mas com absorção fraca no IV: uma vibração νN-Cα, predominante em 1.096 cm^{-1}, e um modo misto a 881 cm^{-1}.

Vibração de cadeias laterais

As cadeias laterais de resíduos de aminoácidos de proteínas absorvem luz no IV. Entretanto, a identificação de resíduos específicos é dificultada para alta sobreposição das suas frequências vibracionais.

Dentre os diferentes grupamentos presentes em cadeias laterais, há dois tipos particulares que absorvem em regiões espectrais livres de sobreposição por outros grupos e que podem, dessa forma, ser assinalados. O primeiro grupamento é a sulfidrila das cisteínas, com absorção entre 2.550 e 2.600 cm^{-1}, e o segundo é a carbonila (C=O) de grupamentos carboxílicos protonados, com absorção entre 1.710 e 1.790 cm^{-1}. A análise destas regiões pode fornecer informações tais como eventos de (des)protonação.

Por exemplo, os resíduos Asp e Glu protonados apresentam duas bandas fortes entre 1.550 e 1.580 cm^{-1} e próximos a 1.400 cm^{-1}. Essas bandas, contudo, são deslocadas na presença de quelantes de cátions dependendo do tipo de coordenação (importante para o estudo de proteínas que ligam íons cálcio).

Por outro lado, a absorção da cadeia lateral de resíduos de Arg, que ocorre em 1.635 e 1.673 cm^{-1}, é sobreposta à absorção da amida I. Contudo, a troca de 1H por 2H gera desvios a -50 e -70 cm-1, respectivamente (desvio para frequências menores), o que permite a visualização destas bandas.

Há ainda uma vibração de Tyr que é frequentemente visualizada em espectros de IV de proteínas a ~1.517 cm^{-1}. Esta frequência vibracional é deslocada para ~1.500 cm^{-1} quando ocorre desprotonação da cadeia lateral do resíduo de Tyr.

11.6. IV e estrutura $2^{ária}$

Como descrito na seção anterior, proteínas apresentam bandas vibracionais características no IV médio. A banda da amida I é a região que fornece informação sobre a estrutura $2^{ária}$ destas macromoléculas.

A frequência exata da primeira vibração (estiramento C=O) depende:

i) da natureza das ligações de hidrogênio que envolvem o grupamento amídico, o que é determinado pela estrutura $2^{ária}$ particular adotada pela proteína;

ii) da orientação e distância dos dipolos que interagem, o que fornece informação sobre arranjo geométrico de grupamentos peptídicos em uma cadeia polipeptídica.

O termo dipolo se refere a dois pólos. Em física, um dipolo elétrico envolve a separação de cargas positivas e negativas (polo positivo e polo negativo). Em moléculas polares, como a água, por exemplo, um dipolo é formado devido a uma distribuição desigual de cargas (elétrons) na ligação covalente (O-H), gerando uma região de carga parcial positiva (hidrogênios) e outra de carga parcial negativa (oxigênios).

Dipolos induzidos são formados quando um íon ou uma molécula dipolar (que apresenta um dipolo permanente) induz a formação de um dipolo em um átomo ou molécula que antes não apresentava uma distribuição de cargas. Quando o oxigênio molecular (O_2, não apresenta um dipolo) interage com uma molécula de água (dipolo permanente), esta última induz um dipolo no O_2.

A aplicação de FTIR para determinação

de conteúdo de estrutura 2ária em proteínas se mostrou viável após a análise experimental do espectro de IV de proteínas com estrutura já resolvida por difração de raios-X, assim como a comparação com outros parâmetros experimentais, como experimentos de dicroismo circular (ver capítulo 10) e cristalográficos (distâncias entre ligações, ângulos de ligação e de diedro). Dessa forma, foi possível estabelecer correlações estruturais-espectrais e, assim, validar a metodologia de FTIR para identificação de componentes de estrutura 2ária em proteínas.

De qualquer modo, é importante ressaltar que não há hoje método capaz de descrever as características conformacionais de proteínas de forma absoluta. Um dos motivos para isto reside na dificuldade em reproduzir, durante os experimentos, as condições do meio nas quais a proteína exerce sua função fisiologicamente, tais como tampão, pH, presença de íons, moduladores, etc, uma vez que as características conformacionais da proteína variam como função destes fatores.

Neste momento, é importante ressaltar que as frequências vibracionais na amida I serão deslocadas para valores menores quando a proteína está diluída em ^2H_2O (Tabela 3-11), o que irá ocorrer quando estamos avaliando estrutura 2ária de proteínas em solução.

A Tabela 3-11 indica as regiões na amida I que são assinaladas aos diferentes componentes de estrutura 2ária. Podemos perceber que há sobreposição entre algumas regiões, o que implica na necessidade de um processamento matemático posterior à coleta do espectro de IV de proteínas, como veremos a seguir. A Figura 12-11 mostra espectros representativos de proteínas ricas em hélices α e em folhas β (vermelho).

Em geral, a vibração das hélices α ocorre a ~1.650 cm^{-1}, e a de estruturas desordenadas a ~1.645 cm^{-1}, proximidade esta que dificulta a avaliação direta do conteúdo de cada um destes componentes na estrutura proteica. As folhas β, por sua vez, apresentam mais de uma região vibracional para a amida I, com bandas de alta (entre 1.670 e 1.690 cm^{-1}) e baixa frequências (de 1.620 a 1.640 cm^{-1}).

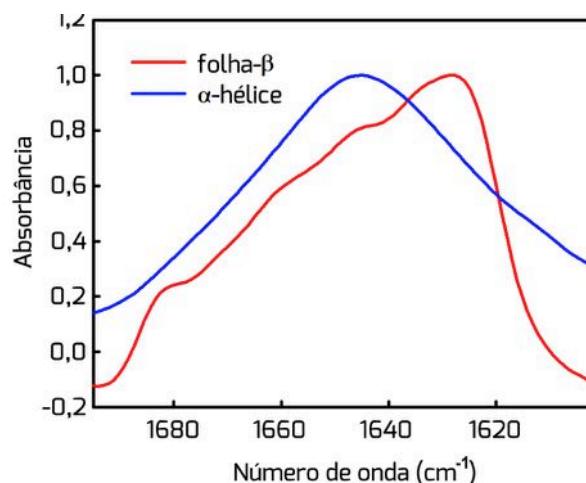



Figura 12-11: Exemplo da região amida I de proteínas com estrutura secundária rica em hélices α (azul) e folhas β (vermelho). Os espectros foram obtidos para proteínas diluídas em ^2H_2O.

Voltas são assinaladas nas regiões de frequência entre 1.660 e 1.680 cm^{-1}.

A análise de folhas β apresenta um desafio particular, pois ainda há incerteza sobre a possibilidade de distinção de folhas β paralelas e antiparalelas por FTIR. O que geralmente se observa é uma separação da amida I em proteínas com alto conteúdo de folhas β antiparalelas. Sendo assim, é possível diferenciar folhas β paralelas de antiparalelas, porque as paralelas absorvem somente em baixos números de onda (banda principal a ~1.630 cm^{-1}) e não possuem o componente em ~1.680 cm^{-1} das folhas β antiparalelas.

Além disso, em alguns casos é possível distinguir entre folhas β antiparalelas intra- e intermoleculares, ou seja, proteínas que formam folhas quando agregadas. Esta agregação promoveria uma absorção em frequências altas (~1.685 cm^{-1}) e baixas (~1.615 cm^{-1}) (Figura 13-11).

11.7. Informações quantitativas

Como vimos acima, há uma grande sobreposição de componentes vibracionais ao longo da banda amida I. Sendo assim, para o assinalamento e quantificação (ou seja, cálculo aproximado da porcentagem dos componentes de estrutura 2ária de uma dada

Tabela 3-11. Assinalamento dos componentes de estrutura secundária de proteínas a partir da análise da amida I. Valores coletados por Byler & Susi (1986) e compilados por Barth & Zcherp (2002).

Estrutura 2ária	Posição do pico na presença de ^1H_2O (cm^{-1})		Posição do pico na presença de ^2H_2O (cm^{-1})	
	Média	Variação	Média	Variação
hélice α	1654	1648 a 1657	1652	1642 a 1660
Folhas β (baixa frequência)	1633	1623 a 1641	1630	1615 a 1639
Folhas β (alta frequência)	1684	1674 a 1695	1675	1671 a 1694
Voltas	1672	1662 a 1686	1671	1660 a 1694
Estruturas desordenadas	1654	1642 a 1657	1645	1639 a 1654

proteína), é necessário realizar um processamento do espectro original, na região desta banda.

Iremos agora abordar como é possível determinar a composição de estrutura 2ária de proteínas a partir da análise da banda amida I (de 1.700 a 1.600 cm^{-1}). Como podemos observar na Figura 14-11, somente com uma inspeção visual da amida I, não é possível identificarmos todos os componentes de estrutura 2ária (com suas diferentes frequências, como mostrado na Tabela 3-11) que formam a

Figura 13-11: Espectro de infravermelho (região amida I) representativo de uma proteína que sofreu agregação induzida por temperatura. As linhas tracejadas indicam componentes de folha β de alta (esquerda) e baixa (direita) frequências.

proteína em questão. Sendo assim, de forma geral, é necessário empregar abordagens matemáticas para separar as frequências vibracionais na banda amida I para o posterior assinalamento dos diferentes componentes (diferentes frequências) de estrutura 2ária. A separação dos diferentes componentes pode ser feita por decomposição da amida I empregando:

i) cálculo da segunda derivada do espectro (Figura 15-11). A largura da banda da derivada assim obtida é menor que a largura da banda original. Assim, a segunda derivada pode ser utilizada para resolver bandas sobrepostas;

ii) realizar uma auto-deconvolução (FSD, *Fourier self-deconvolution*). O princípio de estreitamento de linha da auto-deconvolução é a multiplicação da transformada de Fourier do espectro original por uma função dependente da forma da linha que aumenta com o aumento da distância a partir do pico central. No caso de deconvolução de linhas lorentzianas, se usa uma função exponencial. Dessa forma, as regiões da transformada de Fourier que codificam para estruturas finas no espectro original levam um peso mais forte. Após transformação de volta em um espectro de IV, os componentes do espectro que mudaram mais ao longo do número de onda (ou da frequência) são amplifi-

Figura 14-11: Espectro de IV (região amida I) não processado (vermelho) e após processamento matemático (FSD) da proteína lisozima em ^2H_2O.

Figura 15-11: Espectro na região da amida I de uma proteína em solução (^2H_2O) (vermelho). Em azul está representada a segunda derivada do espectro original e, em preto, o espectro resultante do somatório dos diferentes componentes (verde) deduzidos a partir da segunda derivada.

cados e as bandas então aparecem mais definidas. Para a amplificação, deve-se definir um valor de FWHH (no geral de 13 a 25 cm^{-1}, dependendo da resolução espectral e da relação sinal/ruído) e um fator de incremento, que será multiplicado ao sinal total da amida I;

iii) uma terceira abordagem é de incremento de *fine-structure*; uma versão suavizada do espectro original é multiplicada por um fator pouco menor que 1 e, subsequentemente, subtraída do espectro original, aumentando a estrutura fina do espectro, similarmente a uma FSD.

Existem diversos problemas para a predição de estrutura 2ária por FTIR, independentemente do método aplicado. Não há um único espectro de IV para um tipo de estrutura 2ária, e o espectro obtido também depende de detalhes estruturais como deformações na hélice ou o número de fitas adjacentes em uma folha β. Além disso, outro problema é a absorção por cadeias laterais nesta região. É estimado que de 10 a 30 % da absorção total da amida I é derivada de cadeias laterais.

Após a separação dos diferentes componentes (frequências) da amida I, utilizando alguma das abordagens apresentadas acima, é possível identificar (ver Tabela 3-11) e calcular a fração de cada componente de estrutura 2ária presente na proteína. O percentual de cada tipo de estrutura 2ária é então calculado a partir da área de cada banda correspondente a um determinado tipo de estrutura 2ária em comparação com a área do espectro total na amida I (que apresenta o valor de 100%).

11.8. Desvio de ^1H para ^2H

Como vimos anteriormente, os espectros de IV de proteínas em solução são obtidos a partir de amostras diluídas em ^2H_2O. A troca $^1H/^2H$ leva a pequenos desvios nos componentes da amida I (denominada amida I´ quando a proteína está dissolvida em ^2H_2O). Esses desvios de frequência são causados pela pequena contribuição da dobra N-H para esta banda de vibração.

Para proteínas, a grandeza do desvio depende do tipo de estrutura 2ária. Em geral, ocorre um desvio de ~15 cm^{-1} para componentes de baixa frequência de folhas β e voltas. Estruturas desordenadas sofrem desvio de 10 cm^{-1}, enquanto que para as outras bandas o desvio é menor. A magnitude do desvio vai depender da extensão da contribuição da $\nu_{N\text{-}H}$ para a banda amida I.

Outra causa para este desvio não ser

homogêneo entre todas as proteínas é a troca incompleta de ^1H por ^2H, principalmente em regiões de estrutura 2^{ária} ordenada que apresentam um pequeno desvio. Sendo assim, é essencial o conhecimento do solvente utilizado (se H_2O ou ^2H_2O) para interpretação de espectros de IV de proteínas.

11.9. Vantagens e limitações

Como principais vantagens da técnica, podemos citar:

i) As medidas de FTIR de proteínas podem ser realizadas rapidamente;
ii) Usualmente, os espectros de FTIR apresentam elevada resolução mesmo com sinal baixo;
iii) Pode ser aplicada em amostras em solução ou secas;
iv) Pode ser aplicado a amostras insolúveis, o que usualmente limita as medidas em outras técnicas espectroscópicas;
v) Meios opticamente turvos podem ser utilizados, o que amplia a diversidade de ambientes em que a macromolécula pode ser estudada;
vi) Permite a avaliação da estrutura de proteínas inseridas em membrana e agregados proteicos, além de outros sistemas pouco estudados por outros métodos espectroscópicos;
vii) Grande quantidade de informação obtida;
viii) Técnica não-destrutiva, ou seja, há a possibilidade de recuperação da amostra após a medida.

As limitações e cuidados a serem tomados incluem:

i) A quantidade de proteína necessária é elevada (de 1 a 4 wt%);
ii) A troca $^1H_2O \rightarrow {}^2H_2O$ requer liofilização da amostra;
iii) Avaliação quantitativa ainda limitada devido à falta de modelos acurados;
iv) A deconvolução nem sempre irá representar a estrutura correta final em função do elevado número de bandas sobreponíveis. Amplificação do ruído após FSD.
v) Exige manipulação matemática extensa dos dados experimentais obtidos;
vi) Sofre interferência de contaminantes que absorvam no IV médio, como o TFA, solvente utilizado na purificação de peptídeos sintéticos, que absorve a 1.673 cm^{-1}.

11.10. Conceitos-chave

Caminho óptico: espessura da solução atravessada por um feixe de luz.

Interferograma: Padrão de interferência gerado por um interferômetro, a partir da recombinação da luz gerada a partir de duas fontes diferentes.

FWHH (*full bandwidth at half height*): largura máxima da banda na metade da altura (intensidade total).

Beam splitter: separador do feixe de infravermelho, presente no interferômetro.

FSD: *Fourier self-deconvolution*. Deconvolução de uma região do espectro de IV (Amida I, no caso), a partir de estreitamento de banda e da utilização de um fator de incremento (de 1.5 a 2.5), que é multiplicado pelo sinal da Amida I obtida.

N-metil acetamida (NMA): Menor molécula que contém um grupamento peptídico em trans. Utilizado como modelo para análise dos modos vibracionais da cadeia polipeptídica.

Transformada de Fourier: É uma transformada reversível de uma função em outra função. A segunda função, chamada de transformada de Fourier fornece os coeficientes de funções senoidais (suas frequências) que podem ser recombinadas para obter a função original.

Massa reduzida (μ): Quantidade que permite

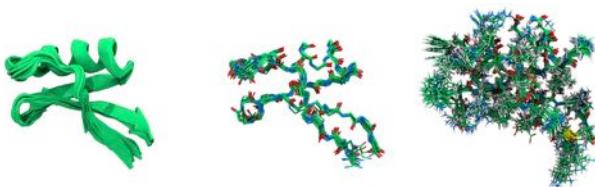
que o problema de dois corpos na mecânica Newtoniana seja resolvido como um problema de um corpo somente, pois:

$\mu = m1 \times m2/m1 + m2$, onde m1 é a massa do corpo 1 e m2 é a massa do corpo 2.

11.11. Leitura recomendada

BARTH, Andreas; ZSCHERP, Christian. What vibrations tell us about proteins. ***Q. Rev. Biophys.*** 35, 369-430, 2002.

BYLER, D. M.; SUSI, H. Examination of the secondary structure of proteins by deconvolved FTIR spectra. ***Biopolymers***. 25, 469-87, 1986.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. Infrared Spectrometry. In: ***Spectrometric identification of organic compounds***. 7a.ed. John Wiley & Sons, 2005.

SUREWICZ, W. K.; MANTSCH, H. H.; CHAPMAN, D. Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment. ***Biochemistry***. 32, 389-94, 1993.

12. Ressonância Magnética Nuclear

Marcius da Silva Almeida

Estrutura 3D da proteína Psd1 determinada por RMN.

12.1. Introdução

Os concomitantes avanços em biologia molecular e em espectroscopia por Ressonância Magnética Nuclear (RMN) multidimensional tiveram como reflexo um aumento explosivo na utilização da espectroscopia por RMN a fim de obter informações estruturais e dinâmicas em macromoléculas biológicas, incluindo ácidos nucleicos, carboidratos e proteínas.

A espectroscopia por RMN em solução e a cristalografia por raios-X são, essencialmente, as únicas técnicas experimentais capazes de fornecer informações da estrutura tridimensional de uma macromolécula com resolução atômica. Aproximadamente 97% das estruturas depositadas no banco de dados *Protein Data Bank* (PDB) resultam da aplicação de uma destas técnicas. As demais estruturas provêm, essencialmente, de modelos teóricos. O número de estruturas resolvidas por cristalografia excede em ~5 vezes as resolvidas por RMN, em grande parte devido a um limite no tamanho da proteína passível de ter sua estrutura determinada por RMN (em torno de 6 kDa por técnicas bidimensionais e ~40 kDa por técnicas de três ou mais dimensões). Em contrapartida, a cristalografia é limitada, principalmente, pela dificuldade na obtenção de monocristais.

A primeira estrutura 3D determinada através de RMN foi do inibidor de α-amilase tendamistat, em 1986, por Kline e colaboradores, ao passo que a primeira estrutura 3D de proteína determinada com alta resolução através de RMN foi da interleucina 1β, em 1991, por Clore e colaboradores.

12.2. Fundamentos

Uma das características de um núcleo atômico é sua rotação em torno do seu próprio eixo, um fenômeno denominado de spin. Os núcleos com spin possuem momento angular p que varia de forma quântica. O número máximo das componentes do momento angular de um núcleo é denominado de número quântico de spin (I). Um núcleo possui $2I+1$ estados de magnetização, onde o componente do magnetismo nuclear possui valores $I, I-1, I-2, ..., -I$.

Em proteínas, os núcleos atômicos mais importantes (devido a propriedades intrínsecas que levam a geração de um sinal plausível de ser identificado por espectroscopia de RMN) são o ^{1}H (abundância natural de 99,98%), o ^{13}C (abundância natural de 1,11%) e o ^{15}N (abundância natural de 0,36%). O número quântico de spin destes núcleos é $1/2$. Desta forma, estes núcleos possuem dois estados de spin ($-1/2$ e $+1/2$).

O spin de núcleos carregados cria um campo magnético orientado paralelamente ao eixo do spin, que pode ser representado por uma quantidade vetorial . Este momento magnético é diretamente proporcional ao momento angular e à constante giromagnética (γ) do núcleo. Consequentemente, os diferentes estados do spin dos núcleos supracitados resultam em dois estados de magnetização, representadas pelo número quântico magnético m, igual à $+1/2$ e $-1/2$.

Para se obter um sinal de RMN destes núcleos em um espectrômetro moderno, inicialmente é induzida a orientação do vetor μ ao longo do vetor de um campo magnético forte gerado por um magneto (vetor B_0). O vetor poderá estar alinhado tanto no mesmo sentido como no sentido contrário ao vetor B_0, sendo que a quantidade de energia que envolve a transição de um núcleo entre estes dois estados é dada pela equação

$$\Delta E = (\gamma \ /2\pi)B_0$$

onde é a constante de Planck.

Como em outras técnicas de espectroscopia, a transição entre estes dois estados pode ser conseguida através da absorção ou da emissão de radiação eletromagnética, em uma frequência v_0 (frequência de Larmor) que corresponde, em energia, à diferença ΔE. Através da equação

$$v_0 = \gamma B_0/2\pi$$

torna-se claro que a frequência da radiação envolvida na transição dos estados energéticos dos spins depende diretamente da força do campo magnético externo e do núcleo estudado. Os espectrômetros de RMN são, em geral, classificados de acordo com a frequência de Larmor do ^{1}H sob a força do campo magnético gerado pelo magneto de tal equipamento. Por exemplo, sob a influência de um campo magnético de 14,1 T, a frequência de Larmor do ^{1}H será de ~600 MHz, e desta forma tem-se um espectrômetro de 600 MHz.

Um fato importante é que os núcleos se distribuem desigualmente entre estes dois estados energéticos, de tal forma que existe um excesso de núcleos no estado de menor energia em relação ao de maior energia. A relação entre o número de núcleos distribuídos entre os dois níveis energéticos é dada pela equação

$$Nj/N_0 = exp(-\gamma \ B_0/2\pi kT)$$

onde N_j é o número de núcleos no estado de maior energia, N_0 é o número de núcleos no estado de menor energia, k é a constante de Boltzmann e T é a temperatura absoluta. No caso do ^{1}H, por exemplo, em um campo magnético de 14,1 T à 293 K, esta relação é de ~0,999901, que significa um excesso de ~198 ppm de prótons no estado de menor energia. Este excesso é representado por um vetor de magnetização resultante M (Figura 1-12).

Através de pulsos de magnetização com vetor perpendicular ao vetor B_0 e na mesma frequência que a frequência de Larmor, é induzida uma reorientação (excitação) do vetor M.

Após certo período de tempo cessa-se o pulso de magnetização e detecta-se o sinal ressonância de cada núcleo enquanto seus vetores M (para cada núcleo) retornam à condição inicial, ou seja, determina-se e frequência de precessão do vetor M de cada núcleo ao passo que estes retomam o alinhamento paralelo com o vetor B_0. Tal fenômeno, que representa o sinal fun-

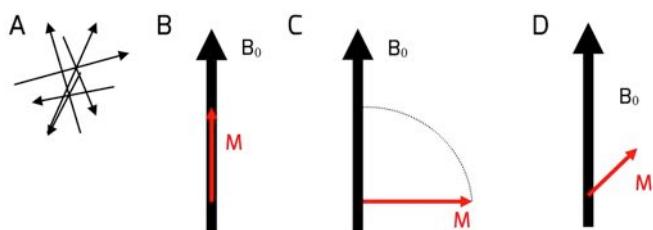



Figura 1-12: Manipulação dos spins para se obter um espectro de RMN. A) inicialmente os núcleos atômicos apresentam vetor de campo magnético μ com orientação caótica. B) Através de um campo magnético forte B_0 é induzida uma orientação coerente dos vetores μ, passando a precessar em torno de B_0. Esta orientação resulta no vetor M (vermelho). C) são gerados pulsos de magnetização perpendiculares ao vetor B_0 com a mesma frequência que da precessão dos spins, o que reorienta o vetor M (que fica perpendicular ao vetor B_0 mas, dependendo da intensidade ou duração do pulso de magnetização, pode ter diversas orientações). D) após os pulsos, ocorre a relaxação (perda de orientação coerente) dos spins, o que é representado pela diminuição do vetor M, assim como seu realinhamento paralelo ao vetor B_0. Neste ultimo momento é realizada a detecção do sinal de ressonância dos núcleos.

damental observado por espectroscopia de RMN, é denominado de *Free-Induction Decay* (FID). Este sinal, representado por uma onda no domínio temporal, é processado, empregando-se o formalismo da transformada de Fourier, e o resultado é um espectro no domínio das frequências.

Nas modernas técnicas de RMN não se utiliza apenas um único pulso de excitação, mas uma sequência de pulsos, que manipulam os spins de uma forma complexa. A manipulação da magnetização dos spins pode revelar influências externas sob um núcleo, como a proximidade ou ligação a outros átomos, através de análises da largura, intensidade e deslocamento químico do sinal de cada núcleo em um espectro de RMN. Desta forma, através destas sequências de pulsos, podem-se obter várias informações relacionadas com a estrutura de uma molécula, que podem por fim, serem "traduzidas" na forma da estrutura tridimensional de uma proteína.

12.3. Deslocamento químico

O deslocamento químico define a localização de uma linha nos espectros de RMN ao longo do eixo de frequência. Esta grandeza é medida relativa a um composto de referência (geralmente um composto solúvel em água como o 3-trimetilsililpropionato).

Nos espectros de RMN a unidade do deslocamento químico de um núcleo é normalmente representada em ppm (partes por milhão), que é uma forma de normalizar todos os espectros em função da intensidade do campo magnético do magneto onde se fez o espectro de uma amostra (como citado anteriormente, a frequência de Larmor depende fortemente da intensidade do campo magnético).

Os núcleos atômicos estão sempre rodeados de diversos átomos e quase sempre estão ligados a outros átomos e, assim, são rodeados por uma nuvem eletrônica. Essa nuvem eletrônica gera campos magnéticos secundários que são os principais responsáveis pela alteração do deslocamento químico de um núcleo em uma macromolécula (efeito denominado de blindagem nuclear).

Através de um espectro de RMN pode-se observar seletivamente o sinal de diferentes núcleos em diferentes ambientes químicos, ou ainda ligados a diferentes átomos. Como exemplificado na Tabela 1-12, no caso da espectroscopia de proteínas por RMN de ^{1}H, podem-se distinguir diversos grupos de átomos de hidrogênio pelo deslocamento químico destes. Assim, o deslocamento químico é um dos mais importantes parâmetros em estudos por RMN.

12.4. Acoplamento escalar

Um dos fatores que influencia na magnetização de um núcleo atômico é a sua ligação com outros átomos. Esta interação é conhecida por acoplamento escalar ou spin-spin, sendo representada pela constante de acoplamento $^{n}J_{ab}$, onde n é o número de ligações covalentes separando os núcleos a e b. Normalmente, o acoplamento escalar se estabelece entre átomos separados por até três ligações químicas.

A constante de acoplamento se mani-

Tabela 1-12: Distinção entre os átomos de hidrogênio dos aminoácidos comuns pelo deslocamento químico (adaptado de Wüthrich, 1986).

Tipo de átomo de hidrogênio	Deslocamento químico (ppm)
CH_3	0,9 – 1,4
CH_2 de V, I, L, E, Q, M, P, R, K	1,6 – 2,3
CH_2 de C, D, N, F, Y, H, W	2,7 – 3,3
CH_2 de S, CH de T e CαH	3,9 – 4,8
Outros CH alifáticos	1,2 – 3,3
CH aromático	6,5 – 7,7
NH de cadeia lateral de N, Q, K, R	6,6 – 7,7
NH da ligação peptídica	8,0 – 8,8
NH indólico	10,2

festa em um espectro de RMN como um pico composto denominado multipleto (sinal dividido em duas ou mais componentes) e sua magnitude é indicada pela distância entre os picos de um multipleto, em hertz (Hz). Um fato importante para a determinação da estrutura de moléculas por RMN é que as constantes de acoplamento $^3J_{ab}$ dependem do ângulo de torção entre os átomos acoplados (maiores detalhes no tópico "cálculo da estrutura").

12.5. Efeito Overhauser nuclear

A influência da magnetização de átomos não ligados por meio de uma ligação química, porém próximos, é o mais importante efeito na magnetização de um núcleo para a determinação da estrutura de proteínas por RMN.

Tal fenômeno, denominado de efeito Overhauser nuclear (NOE), ocorre devido ao acoplamento dipolar (pelo espaço) entre diferentes núcleos, que envolve a transferência de magnetização entre os spins acoplados.

A intensidade do acoplamento dipolar é proporcional ao inverso da sexta potência da distância entre os átomos, sendo que este tipo de interação é normalmente detectado entre átomos distantes entre si em até 5 Å.

12.6. Estrutura de proteínas

Um dos passos para se determinar a estrutura tridimensional de macromoléculas por espectroscopia de RMN é o assinalamento (identificação) dos picos de ressonância.

Em proteínas, devido à grande quantidade de átomos, ocorre uma enorme sobreposição de sinais nos espectros de RMN, o que torna impraticável o assinalamento dos picos de ressonância. Uma forma de resolver este problema é a utilização de espectroscopia bidimensional, através de uma série de sequências de pulsos específicas.

Os espectros bidimensionais essenciais para a determinação da estrutura de proteínas incluem o TOCSY e o NOESY, ambos de correlação homonuclear. Espectros de correlação heteronuclear podem ser incluídos no processo de determinação da estrutura de proteínas, de forma a facilitar o trabalho de assinalamento dos sinais de ressonância. Neste caso estão incluídos o HMQC (*heteronuclear multiple quantum coherence*) e HSQC (*heteronuclear single quantum coherence*).

TOCSY

Também conhecido por HOHAHA (*Homonuclear Hartmann-Hahn*), o experimento de TOCSY (*Total Correlated Spectroscopy*) consiste em uma sequência de pulsos que induzem a transferência da magnetização entre núcleos, como prótons ou carbono, via acoplamento escalar.

Uma vez que a transferência via acoplamento escalar por mais de quatro ligações é praticamente nula e que o carbono da carbonila da ligação peptídica não possui próton ligado, o 2D [^{1}H,^{1}H]-TOCSY de proteínas evidenciará interação somente entre prótons de cada aminoácido isoladamente (Figura 2-12). Neste espectro, cada pico (denominados de picos de correlação e representados por curvas de nível) indica a presença da interação entre dois prótons via acoplamento escalar. O conjunto dos sinais de correlação dos prótons de um aminoácido é denominado de sistema de spin (Figura 2-12).

NOESY

O NOESY (*Nucear Overhauser Effect SpectroscopY*) é o espectro crucial para a determinação da estrutura de uma proteína. Neste tipo de experimento é induzida, através de uma sequência de pulso específica, a transferência de magnetização entre os núcleos via acoplamento dipolar, que depende da proximidade entre átomos, mesmo que não estejam ligados quimicamente.

No espectro de 2D [^{1}H,^{1}H]-NOESY aparecerão sinais (os NOEs) que representam prótons próximos (distância menor do que 5 Å). A intensidade dos NOEs depende de vários fatores, dentre eles, a distância entre os prótons acoplados (o que por uma aproximação simplista, pode representar diferentes limites

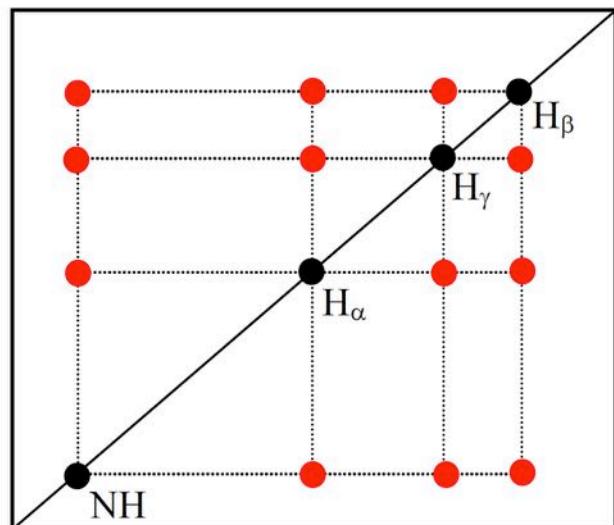
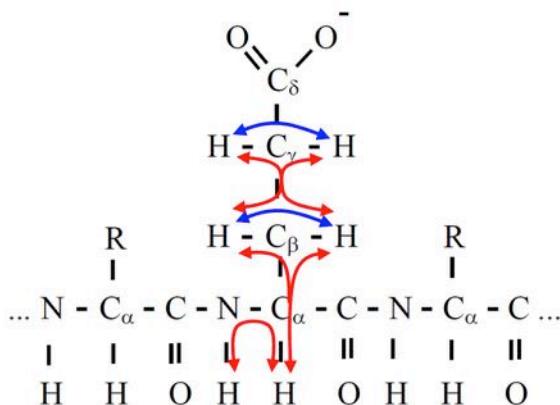



Figura 2-12: Prótons em acoplamento escalar de um fragmento peptídico contendo o resíduo de ácido glutâmico. As setas vermelhas e azuis indicam acoplamento ${}^{3}J_{HH}$ e ${}^{2}J_{HH}$, respectivamente, que só podem ocorrer entre prótons de um mesmo aminoácido. A cadeia lateral dos aminoácidos que precedem e sucedem o glutamato é representada pela letra R. Ao lado do fragmento polipeptídico é apresentado o desenho esquemático de um espectro bidimensional de $[{}^{1}H,{}^{1}H]$-TOCSY, evidenciando os picos de correlação dos prótons do glutamato (círculos vermelhos). Tal perfil representa um sistema de spin. Os picos em preto na diagonal do espectro são os sinais de ressonância de cada próton do glutamato.

de distância entre os prótons). Desta forma, pode-se fazer uma aproximação semi-quantitativa entre a intensidade dos NOEs e a distância que separa os prótons acoplados. NOEs intensos representam prótons separados por 1,8 a 2,7 Å, NOEs de intensidade média representam prótons separados por 1,8 a 3,4 Å e NOEs fracos, prótons separados por 1,8 a 5,0 Å.

Usando-se os dados de distância entre prótons de uma proteína indicadas pelos NOEs (restrições de NOE), pode-se finalmente criar um modelo estrutural desta macromolécula.

Espectros 2D heteronucleares

Nos experimentos bidimensionais heteronucleares (HMQC - *Heteronuclear Multiple Quantum Coherence* ou então HSQC - *Heteronuclear Single Quantum Coherence*), é realizada a transferência de magnetização entre o spin do próton e o spin de outro núcleo atômico, através de somente uma ligação química. Nos espectros aparecerão picos de correlação entre próton e ${}^{13}C$ ou então entre próton e ${}^{15}N$, sendo este sinal importante para a caracterização geral da conformação da proteína, assim como da qualidade da amostra a ser estudada. Na Figura 3-12 são exemplificados dois espectros, um de uma proteína bem enovelada e estável, sendo por isso passível de ter sua estrutura determinada por RMN em solução, assim como um de uma proteína desordenada e agregada.

Espectros de tripla ressonância

Em experimentos de tripla ressonância pode-se associar a magnetização entre diferentes núcleos para obter um mapeamento bem definido dos sinais de uma proteína. Como por exemplo, com o espectro tridimensional (3D) de HNCO, ter-se-á um sinal oriundo da transferência de magnetização entre próton amídico, nitrogênio amídico e carbono da carbonila (Figura 4-12). Espectros mais complexos geralmente são usados para obter correlações entre os diversos núcleos de uma proteína e, assim, conseguir uma descrição (assinalamento) o mais completa possível da cadeia polipeptídica. Por exemplo, pode-se

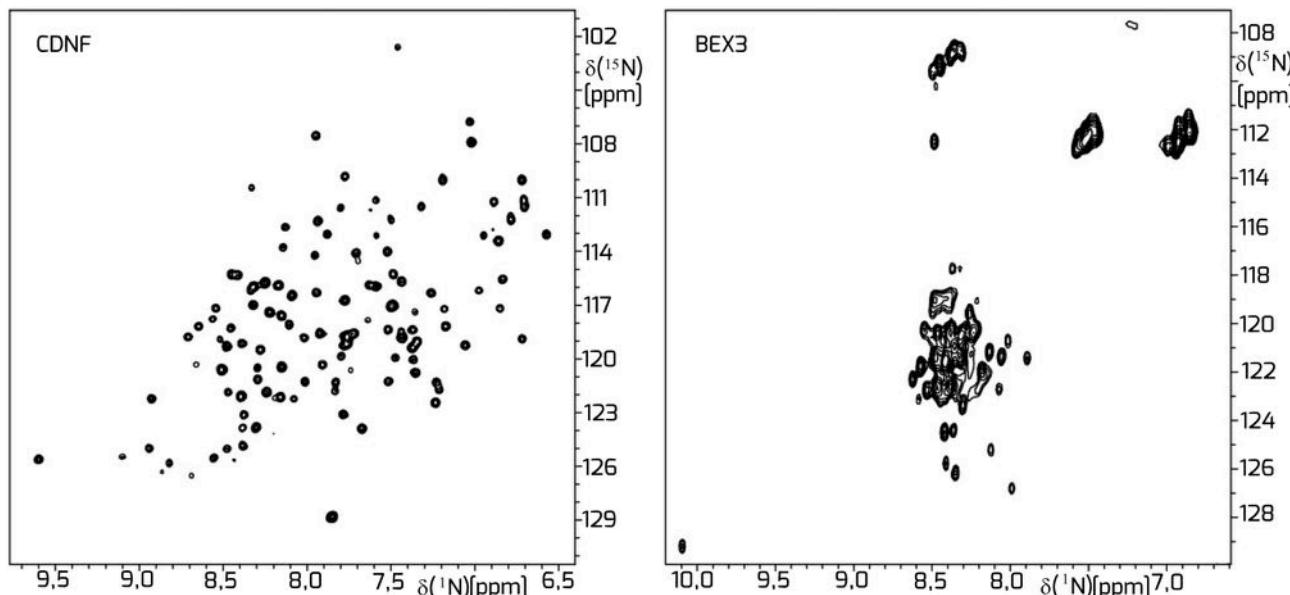



Figura 3-12: Espectros bidimensionais heteronucleares 2D [^{1}H,^{15}N]-HSQC de duas proteínas, CDNF (*Cerebral Dopamine Neurotrophic Factor*, contendo 162 resíduos de aminoácidos ou 18,4 kDa) e BEX3 (*Brain Expressed X-linked*, contendo 124 resíduos de aminoácidos ou 14,5 kDa), que representam uma proteína bem enovelada e uma proteína com alto grau de desordem, respectivamente. O espectro de uma proteína bem enovelada apresenta diversos sinais bem dispersos e bem definidos, diferente do espectro de uma proteína desenovelada e com grande tendência de agregação, que exibe picos sobrepostos. Proteínas com caraterísticas espectrais similares a CDNF normalmente podem ter suas estruturas determinadas por RMN.

fazer um espectro 6D HNCOCANH, onde a magnetização será transferida entre hidrogênios amídicos de aminoácidos vizinhos através da carbonila e do carbono alfa (Figura 5-12).

Para se determinar os sistemas de spin de uma proteína, são necessários pelo menos quatro espectros de tripla ressonância (3D HNCO, 3D HNCACB, 3D CBCA(CO)NH e 3D HBHA(CO)NH) e dois espectros tridimensionais de TOCSY, um editado para ^{13}C e outro para ^{15}N. Alternativamente, é possível usar métodos ainda mais modernos de determinação de estruturas de proteínas por RMN, tais como a aquisição de dois espectros de quatro dimensões (4D e 4D) e um de cinco dimensões (5D).

Em qualquer caso, a análise destes espectros de tripla ressonância deve ser complementada por espectros de [^{1}H,^{1}H]-NOESY tridimensionais, editados para ^{13}C e ^{15}N que evidencia NOEs entre prótons, desde que um deles esteja ligado a um ^{13}C ou ^{15}N, respectivamente.

12.7. Análise dos espectros de RMN

Para qualquer estudo de proteínas por espectroscopia de RMN, cada sinal de ressonância deve ser associado a um núcleo específico. Este processo é denominado de atribuição das ressonâncias.

A atribuição das ressonâncias de uma proteína é obtida através da análise em conjunto dos espectros de NOESY, TOCSY, espectros bidimensionais heteronucleares e de tripla ressonância, onde o intuito é correlacionar cada um dos sinais de ressonância encontrados nestes espectros com os prótons, carbonos e nitrogênios de cada um dos aminoácidos da proteína. Neste processo, inicialmente as ressonâncias de ^{1}H, ^{13}C e ^{15}N são classificadas (em HN, Hα, Hβ, Cα, Cβ e CO, dentre outros) de acordo com seus deslocamentos químicos (Figura 6-12 e 7-12).

Os espectros são então analisados por regiões de acordo com o tipo de grupamento químico esperado em cada faixa de deslocamento químico. Apesar de serem observadas diferenças entre os sistemas de spin de cada

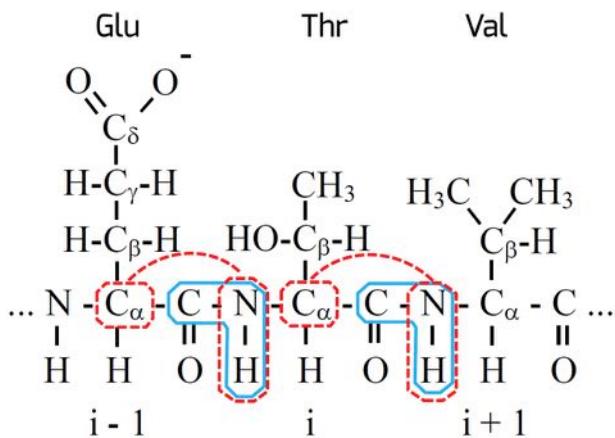



Figura 4-12: Segmento tripeptídico de uma proteína hipotética com a indicação de alguns caminhos de transferência de magnetização obtidos através de dois experimentos de tripla ressonância (3D HNCO em azul e 3D HN(CO)CA em vermelho). O sinal observado conterá informações de deslocamento químico de cada um dos átomos indicados, em um espectro de três dimensões (^{1}H, ^{13}C e ^{15}N).

Figura 5-12: Segmento tripeptídico de uma proteína hipotética com a indicação do caminho de transferência de magnetização obtido por um experimento de tripla ressonância 6D HNCOCANH. O sinal observado conterá informações de deslocamento químico de cada um dos átomos indicados, em um espectro de seis dimensões (^{1}H, ^{15}N, ^{13}CO, ^{13}Cα, ^{15}N e ^{1}H). Note que este tipo de espectro identifica a ligação de um sistema de spin (aminoácido) a outro.

aminoácido, podem ocorrer sobreposições de picos de correlação nos espectros e alterações acentuadas de deslocamento químico de um núcleo atômico (em uma proteína bem estruturada, cada próton poderá estar localizado em um ambiente químico particular e, por isto, sofrer diferentes graus de blindagem nuclear), o que dificulta a análise dos espectros de RMN no que diz respeito à identificação dos sistemas de spin.

Em vista disto, o passo seguinte é identificar alguns sistemas de spin bem característicos nos espectros, levando-se em conta os valores de deslocamento químico médio dos prótons dos ^{13}C e dos ^{15}N aos quais os prótons estão ligados, das diversas proteínas já estudadas por RMN (por exemplo, note a diferença entre os deslocamentos químicos da alanina e glicina, Figura 7-12).

Os aminoácidos com deslocamento químico de ^{1}H e ^{13}C mais característicos são:

i) glicina, que possui dois Hα ligados a um Cα com deslocamento químico anormalmente baixo (~45 ppm contra ~60 ppm dos outros Cα do restante dos aminoácidos);

ii) treonina, que possui um único Hβ com deslocamento químico anormalmente alto (~4 ppm contra os ~2,5 ppm dos Hβ dos outros aminoácidos) e uma metila com deslocamento químico de Hγ em ~1,5 ppm e intensidade de sinal alta;

iii) serina que possui dois Hβ com deslocamento químico anormalmente alto (~4 ppm);

iv) alanina, que possui uma $C\beta H_3$ que resulta em um pico de Hβ intenso com deslocamento químico em ~1,39 ppm.

A partir da identificação destes aminoácidos bem característicos, nos espectros, buscam-se conectividades entre os sistemas de spin usando diversos espectros.

Em se tratando de proteínas que não estão isotopicamente enriquecidas com ^{13}C e ^{15}N, se usa espectros de NOESY para este processo de atribuição das ressonâncias associados à sequência de aminoácidos da proteína em estudo. O objetivo é buscar NOEs entre prótons da cadeia principal de aminoácidos vizinhos que estão quase sempre a me-

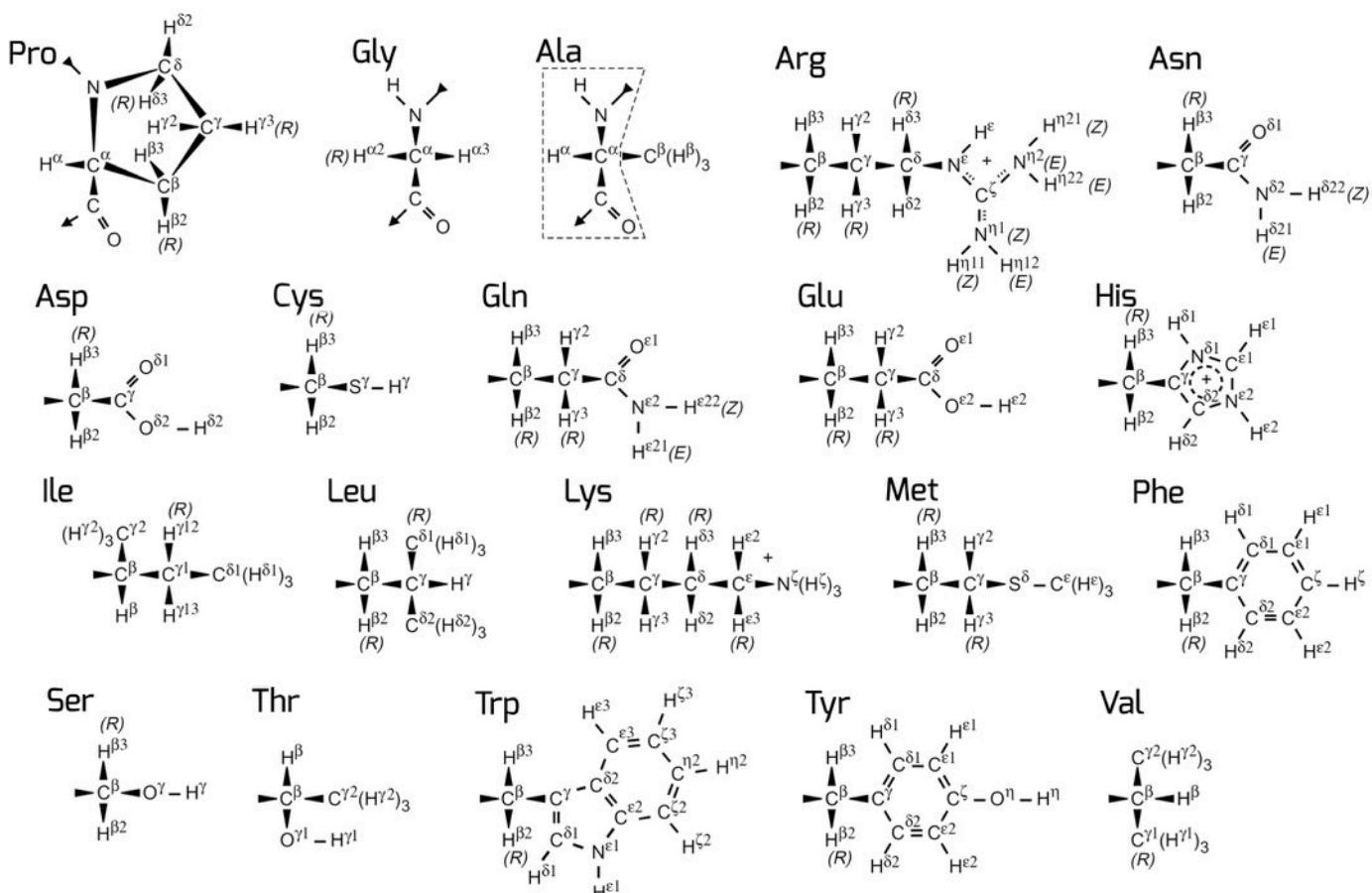



Figura 6-12: Estrutura dos 20 aminoácidos naturais encontrados em proteínas. As nomenclaturas oficiais de cada átomo são evidenciadas. A porção referente à cadeia principal só é representada para a Pro, Gly e Ala. Para todos os outros aminoácidos, a cadeia principal é idêntica à da Ala, que está circulada por uma linha tracejada. Figura extraída com permissão do artigo "Recommendations for the presentation of NMR structures of proteins and nucleic acids (IUPAC© Recommendations 1998)" escrito por Markley e cols. 1998.

nos de 5 Å de distância entre si (Figura 7-12). Neste sentido, procura-se conectividade do HN do aminoácido com sistema de spin atípico identificado (na posição i da sequência da proteína) com Hα, HN e, algumas vezes, Hβ do aminoácido que o precede na sequência polipeptídica (posição $i - 1$), assim como conectividades do Hα, HN e às vezes do Hβ do aminoácido identificado (posição i) com o HN do aminoácido que o sucede (posição $i + 1$) (Figura 8-12).

Em se tratando de proteínas isotopicamente enriquecidas com ^{13}C e ^{15}N (obtidas normalmente quando produzidas em bactérias como *Escherichia coli* ou leveduras como *Pichia pastoris* crescidas em meios sintéticos contendo $^{15}NH_4Cl$ como única fonte de nitrogênio e ^{13}C-Glicose ou ^{13}C-metanol como únicas fontes de carbono), as conectividades entre sistemas de spin também usualmente são obtidas por intermédio de interações escalares, evidenciadas pelos espectros de tripla ressonância discutidos acima (exemplificados nas Figuras 4-12 e 5-12). Desta forma, ou usando espectros de NOESY, obtêm-se algumas sequências tripeptídicas atribuídas ao longo da sequência polipeptídica da proteína.

A partir daí continua-se a atribuição sequencial, levando em conta algumas características dos aminoácidos menos atípicos em conjunto com a sequência 1ária da proteína. A seguir são descritas brevemente algumas peculiaridades dos aminoácidos menos atípicos:

v) valina, só possui um Hβ e dois $C\gamma H_3$ com pico de Hγ com intensidade relativamente alta;

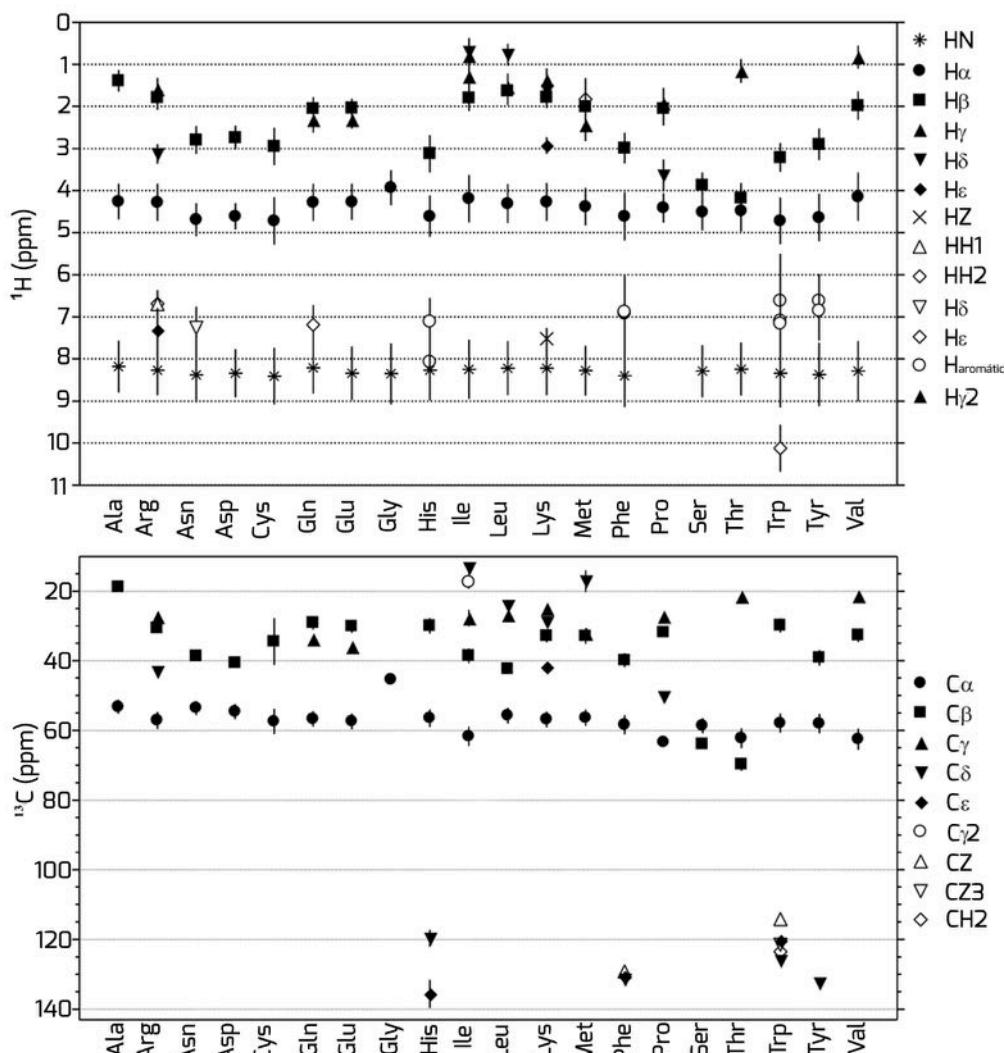



Figura 7-12: Deslocamento químico de ¹H e ¹³C (em ppm) dos átomos dos 20 aminoácidos naturais encontrados em proteínas. As nomenclaturas oficiais de cada átomo são representadas por diferentes símbolos. Valores obtidos do "Biological Magnetic Resonance Data Bank" (http://www.bmrb.wisc.edu). As barras representam os desvios padrões.

vi) leucina, possui longa cadeia lateral, o que pode resultar em uma faixa de sinais de ¹H com deslocamento químico baixo (~1,5 ppm);
vii) isoleucina, apresenta padrão muito semelhante ao da leucina, porém ao contrário da outra, só possui um Hβ;
viii) cisteína e aspartato, suas cadeias laterais se restringem a dois Hβ;
ix) asparagina, através do espectro de NOESY identifica-se conexão entre HN, Hα e Hβ com os prótons amídicos da cadeia lateral (Hδ21 e Hδ22), diferenciando-a da cisteína e do aspartato;
x) histidina, pelo espectro de NOESY é possível ver conectividade entre HN, Hα e Hβ com Hδ2 do anel aromático;
xi) tirosina e fenilalanina, apresentam NOE entre Hβ e Hδ do anel aromático;
xii) triptofano, identificável por NOEs entre Hβ e os Hδ1 e Hε2 do anel aromático (o último próton possui deslocamento químico atípico de ~10 ppm);
xiii) metionina, o intenso pico metílico Hε é facilmente identificável e sua correlação com o resto do sistema de spin se dá somente via NOEs;
xiv) glutamato, possui dois CH_2 na cadeia lateral;
xv) glutamina, além dos dois CH_2 possui conectividade via NOE entre Hγ e prótons amídicos Hε21 e Hε22 da cadeia lateral;
xvi) arginina, identificável através dos

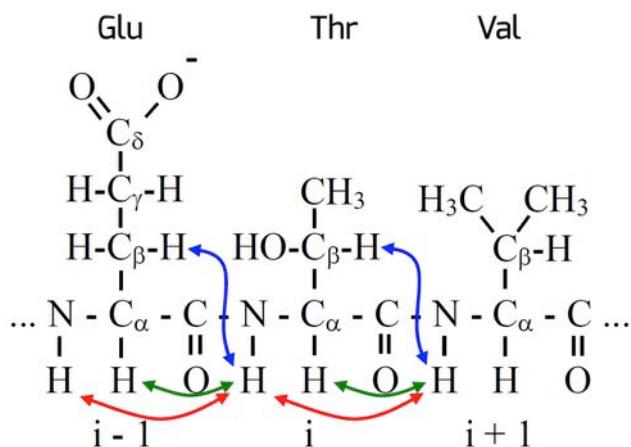



Figura 8-12: Segmento tripeptídico de uma proteína hipotética com a indicação dos NOEs sequenciais empregados para atribuir as ressonâncias dos três sistemas de spin a partir da treonina (resíduo na posição i da sequência da proteína). As cores das setas representam os tipos de NOEs sequenciais (azul - dβN; verde - dαN; vermelho – dNN). A probabilidade de uma conectividade ser realmente sequencial é de aproximadamente 66-79% para NOEs tipo dβN, 76-94% para dNN e 72-98 % para dαN. Quando encontradas duas destas conectividades, a probabilidade delas representarem dois resíduos consecutivos é de 90-99%.

picos de correlação entre CH_2 e NεH da cadeia lateral em adição aos picos entre HN, Hα e os CH_2 da cadeia lateral;

xvii) lisina, como a leucina e a isoleucina possui longa cadeia lateral, o que pode resultar em uma faixa de sinais de ^{1}H, porém com deslocamento químico entre ~1,5 e 3,0 ppm. Além disso, ao contrário dos outros dois aminoácidos, este possui apenas CH_2 na cadeia lateral.

Uma vez tendo todos ou quase todos sistemas de spin identificados (normalmente chega-se ao ponto de identificar 95% dos sistemas de spin), segue-se com o cálculo da estrutura 3D baseada nestes sistemas e nos sinais a serem identificados nos espectros de NOESY e transformados em restrição de distância pelo programa de cálculo de estrutura.

No decorrer do cálculo da estrutura, pode-se identificar mais facilmente possíveis atribuições erradas pelo aparecimento súbito de grandes violações de NOEs (restrições de distância impostas por NOEs que não conseguem ser ajustadas em uma estrutura tridimensional calculada). Uma vez constatado um erro de atribuição, retorna-se ao passo de assinalamento sequencial, trocando-se os sistemas de spin atribuídos erroneamente.

12.8. Cálculo da estrutura

A determinação da estrutura tridimensional de macromoléculas por RMN é baseada, principalmente, em informações de distâncias interprótons (os NOEs). Como citado anteriormente, através das intensidades dos NOEs pode-se fazer uma aproximação da distância entre prótons envolvidos em acoplamento dipolar, distância esta que varia de 1,8 – 5 Å.

Informações adicionais, como ângulos torcionais, podem ser bastante úteis na determinação da estrutura tridimensional de uma proteína, restringindo mais ainda o espaço conformacional adotado pelas estruturas tridimensionais calculadas. Os ângulos de diedro φ (formado pelas ligações entre C_i-$C\alpha_i$ e C_{i-1}-N_i ao longo da ligação entre $C\alpha_i$-N_i do resíduo i) e χ^1 (formado pelas ligações entre N_i-$C\alpha_i$ e $X\gamma_i$-$C\beta i$ ao longo da ligação entre $C\alpha_i$-$C\beta_i$ do resíduo i, onde X pode ser O, C ou S) podem ser inferidos via constante de acoplamento $^3J_{HNH\alpha}$ e $^3J_{H\alpha H\beta}$, respectivamente (Figura 9-12).

A constante de acoplamento $^3J_{HNH\alpha}$ é indicada através da distância entre os picos do dubleto associado à correlação entre HN e Hα, em Hz. A partir daí convenciona-se que para $^3J_{HNH\alpha}$ > 8 Hz tem-se um ângulo φ de aproximadamente -140° e para $^3J_{HNH\alpha}$ < 6 Hz tem-se ângulo φ de aproximadamente -60° (estes ângulos são característicos para segmentos peptídicos em conformação de fita β e hélice α, respectivamente).

Os ângulos de diedro φ e ψ (este último formado pelas ligações entre N_{i+1}-C_i e $C\alpha_i$-N_i ao longo da ligação entre $C\alpha_i$-C_i do resíduo i) podem ser inferidos a partir do índice de deslocamento químico dos núcleos (CSI), uma vez

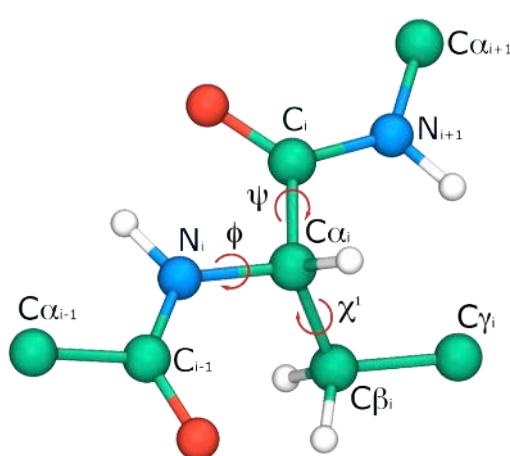



Figura 9-12: Fragmento de uma cadeia polipeptídica evidenciando os ângulos de diedro φ, ψ e χ^1. As linhas pontilhadas indicam as ligações às quais tais ângulos torcionais se referem. As setas vermelhas indicam a rotação das ligações que representam os vértices destes ângulos.

que o deslocamento químico de um núcleo é sensível ao ambiente e a geometria das ligações químicas.

Os deslocamentos químicos de $^{13}C\alpha$ e $^{1}H\alpha$ são os mais usados e melhor correlacionados com a presença de estruturas 2^{árias} em proteínas. Quando o deslocamento químico do $^{13}C\alpha$ de uma série de pelo menos quatro aminoácidos está aumentado em relação aos valores médios oriundos de diversas estruturas proteicas (CSI +), é sugerida a presença de um segmento em hélice α, com ângulos de diedro φ e ψ próximos de -120º e -60º, respectivamente. No caso contrário, quando o deslocamento químico do $^{13}C\alpha$ de uma série de pelo menos quatro aminoácidos está diminuído (CSI -), é sugerido a presença de um segmento em fita β, com ângulos de diedro φ e ψ próximos de -120º e 120º, respectivamente.

No caso do deslocamento químico do $^{1}H\alpha$ o inverso ocorre, ou seja, quando seu deslocamento químico em uma série de pelo menos quatro aminoácidos está acima dos valores médios (CSI +), é indicação de fita β e quando está abaixo de um valor teórico (CSI -), é sugerida a presença de um segmento em hélice α. O cálculo do CSI, sigla para *Chemical Shift Index*, pode ser feito através do endereço eletrônico www.bionmr.ualberta.ca/bds/-software/csi/latest/csi.html.

A análise do ângulo χ^1 fornece importante informação sobre a conformação da cadeia lateral dos aminoácidos, permitindo inclusive o assinalamento estéreo-específico dos dois prótons Hβ (Tabela 2-12). A partir da rotação χ^1 ao redor da ligação Cα-Cβ, as configurações energicamente mais favoráveis são aquelas com o ângulo χ^1 de 60º, 180º ou –60º. Como apresentado na Tabela 2-12, a identificação dos rotâmeros e o assinalamento estéreo-específico dos Hβ se dá a partir da identificação das constantes de acoplamento

Tabela 2-12: Caracterização dos três rotâmeros possíveis em torno da ligação Cα-Cβ. As orientações gauche e trans são referidas como *g* e *t*, onde os índices 2 e 3 indicam os prótons Hβ2 e Hβ3. O padrão de intensidade dos NOEs é indicado para cada conformação.

Características	Conformação		
	g^2g^3	g^2t^3	t^2g^3
	R¹, HN, CO, Cα, Hβ3, Hβ2, Hα	Hβ3, HN, CO, Cα, Hβ2, R¹, Hα	Hβ2, HN, CO, Cα, R¹, Hβ3, Hα
χ^1	60º	180º	-60º
$^3J_{H\alpha H\beta 2}$ (Hz)	2,6-5,1	2,6-5,1	11,8-14,0
$^3J_{H\alpha H\beta 3}$ (Hz)	2,6-5,1	11,8-14,0	2,6-5,1
NOE (Hα, Hβ2)	Forte	Forte	Fraco
NOE (Hα, Hβ3)	Forte	Fraco	Forte
NOE (HN, Hβ2)	Fraco	Forte-médio	Forte
NOE (HN, Hβ3)	Forte-médio	Forte	Fraco

$^3J_{H\alpha H\beta 2}$ e $^3J_{H\alpha H\beta 3}$ e dos NOEs $d_{H\beta 1HN}$, $d_{H\beta 2HN}$, $d_{H\beta 1H\alpha}$ e $d_{H\beta 2H\alpha}$.

Adicionalmente, a informação de ligação de hidrogênio inferida a partir da taxa de troca de próton amídico pode ser agregada ao cálculo da estrutura. As estruturas 2árias regulares estabilizadas por ligações de hidrogênio "protegem" os prótons amídicos envolvidos nestas estruturas, o que se caracteriza por uma baixa taxa de troca destes por prótons do solvente.

Para evidenciar tais prótons "protegidos", dissolve-se a amostra a ser analisada em ^2H_2O e faz-se um espectro bidimensional de ^1H. Se o próton da molécula analisada não estiver "protegido" ele trocará quase que imediatamente por deutério, proveniente da ^2H_2O, desaparecendo seu sinal nos espectros de ^1H-RMN. (o deutério possui frequência de ressonância bem distinta do seu isótopo). A identificação dos prótons com baixa taxa de troca por deutério permite usar restrições estruturais de pontes de hidrogênio no cálculo da estrutura da macromolécula em estudo.

As restrições de distância obtidas por NOEs, assim como de distância entre prótons envolvidos em ligações de hidrogênio inferidas pela taxa de troca de hidrogênio por deutério e as restrições de ângulos φ, ψ e χ^1, inferidas pelas constantes de acoplamento e CSI, são então usadas em protocolos de dinâmica molecular realizados por programas específicos para ajustar a estrutura da proteína a estas restrições, levando em conta a obediência à geometria ideal de ângulos e comprimento de ligações químicas e dos raios de van der Waals dos átomos.

Nestes programas, as moléculas são inicialmente submetidas a uma condição de alta energia cinética (temperaturas de ~50.000 K). Nesta situação, as moléculas estão totalmente desprovidas de qualquer estrutura tridimensional predominante, porém já agregam parâmetros estruturais providos por restrições empíricas (determinadas por um campo de força). Gradualmente, é decrescida a temperatura do sistema (geralmente até 0 K), ao passo que são adicionadas as restrições experimentais.

Através deste procedimento, o programa busca conformações da molécula que satisfaçam o máximo possível às restrições empíricas e experimentais. Finalmente é permitida uma "relaxação" da molécula (passo de minimização e refinamento estrutural) em uma temperatura ainda baixa, porém sob menor influência das restrições de NOE e de van der Waals, de forma a corrigir pequenas imperfeições conformacionais da estrutura como ligações excessivamente torcidas. Neste passo final, a "relaxação" da estrutura é evidenciada pela diminuição da energia do sistema (energias diretamente relacionadas com o grau e número de violações das restrições empíricas e experimentais).

Estes passos são repetidos várias vezes, de forma a obter um conjunto de estruturas (normalmente em torno de 20 estruturas) que são avaliadas, com auxílio de programas, quanto à existência de conformações impróprias ou improváveis. Esta família de estruturas determinadas por espectroscopia de RMN representa uma estrutura tridimensional com pequena variação do espaço conformacional, que é representada por cada uma das estruturas calculadas (exemplo na Figura 10-12). Estruturas com alta resolução obtidas por RMN geralmente possuem um desvio dos átomos da cadeia principal da proteína em relação a uma estrutura média de aproximadamente 0,6 Å.

12.9. Conceitos-chave

Constante de Boltzmann: é uma constante que relaciona energia, no nível de partícula individual, com temperatura. Tem um valor aproximado de $1{,}3806 \times 10^{-23}$ J/K.

Constante de Planck: é uma constante de proporcionalidade entre energia e frequência. Tem um valor aproximado de $6{,}6261 \times 10^{-34}$ J.s.

Constante giromagnética: é a razão entre o momento de dipolo magnético e o momento angular, sendo representada normalmente pelo símbolo gama (γ). Cada

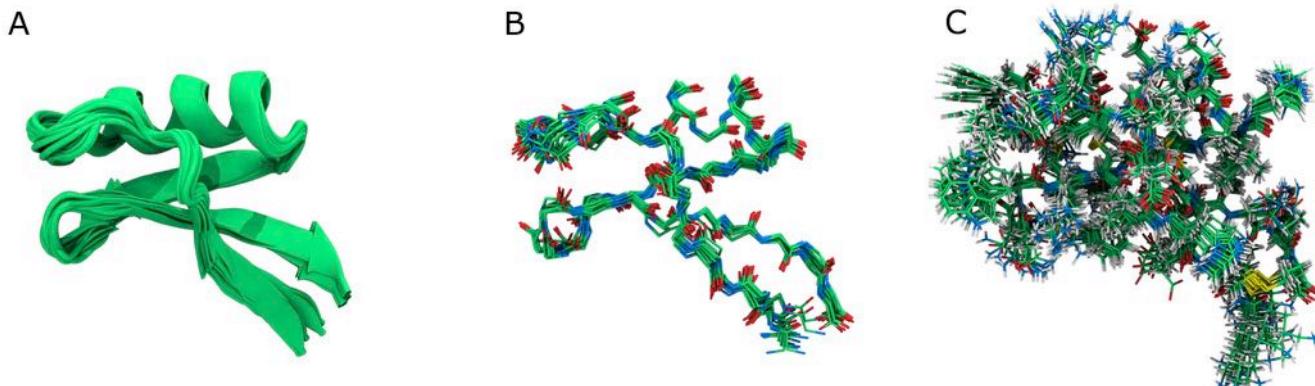



Figura 10-12: Estrutura 3D da proteína Psd1 determinada por RMN. Nesta figura é mostrada uma sobreposição de vinte estruturas obtidas como descrito acima, usando proteína nativa, não enriquecida isotopicamente. Em A, um desenho evidenciando as estruturas secundárias. Em B, são mostrados apenas os átomos da cadeia principal (verde – carbono, azul – nitrogênio e vermelho – oxigênio). Em C, são mostrados todos os átomos (cinza – hidrogênio e amarelo – enxofre). As estruturas estão com o mesmo alinhamento.

núcleo atômico possui uma constante giromagnética específica, sendo a principal razão para que cada núcleo atômico possua uma frequência de RMN distinta em um mesmo campo magnético externo.

Correlação heteronuclear: se diz quando é conseguida uma relação entre núcleos de tipos distintos de átomos em uma molécula. Pode ser correlação escalar ou dipolar, ou seja, dependente ou não dos átomos estarem associados por intermédio de ligações químicas.

Correlação homonuclear: se diz quando é conseguida uma relação entre núcleos do mesmo tipo atômico em uma molécula. Pode ser correlação escalar ou dipolar, ou seja, dependente ou não dos átomos estarem associados por intermédio de ligações químicas.

Projeções de Newman: forma de representação de moléculas que evidencia conformações em relação a uma ligação carbono-carbono tida como referência. O carbono proximal é representado como um ponto e o distal como um círculo (ver Tabela 2-12).

Rotâmero: é uma molécula isomérica em relação à rotação ao redor de uma ligação química simples, normalmente entre carbonos com configuração de orbital de valência tipo sp^3.

Spin: em mecânica quântica e física de partículas, spin é uma forma de momento angular intrínseca de partículas elementares, incluindo o núcleo atômico, quando aplicada para RMN. Em uma das formas de representação, o spin é uma quantidade vetorial com magnitude e direção. O spin nuclear é identificado pelo número quântico de spin e para existir o efeito de RMN o spin deve ser diferente de zero, condição alcançada quando o numero de prótons e/ou nêutrons é ímpar.

Transformada de Fourier: é uma manipulação matemática normalmente usada para transformar funções temporais $f(t)$, em uma função de frequência, cuja unidade geralmente é Hertz.

14.10. Leitura recomendada

ALMEIDA, M. S.; et al. Solution structure of Pisum sativum defensin 1 by high resolution NMR: plant defensins, identical backbone with different mechanisms of action. ***J. Mol Biol.*** 315, 749-57, 2002.

SERRANO, P.; et al. The J-UNIO protocol for automated protein structure determination

by NMR in solution. ***J. Biomol NMR***. 53, 341-354, 2012.

KAY, L. E.; et al. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. ***J. Mag. Res.*** 89, 496–514, 1990.

MARKLEY, J. L.; et al. Recommendations for the presentation of NMR structures of proteins and nucleic acids. ***Pure Appl. Chem.***, 70, 117-142, 1998.

WISHART, D. S.; SYKES, B. D.; RICHARDS, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. ***J. Mol. Biol***. 222, 311–333, 1991.

WÜRTHRICH, K. ***NMR of Proteins and Nucleic Acids***. New York: Wiley, 1986.

13. Cristalografia de Proteínas

Topologia geral dos receptores acoplados à proteína G.

Fernando V. Maluf
João Renato C. Muniz
Glaucius Oliva
Rafael V. C. Guido

13.1. Introdução

A cristalografia de raios-X é uma ciência dedicada ao estudo da estrutura molecular e cristalina, bem como das relações entre essa estrutura e suas propriedades. A cristalografia de raios-X moderna apresenta aplicações amplas nas ciências dos materiais, química, mineralogia, física, matemática e biologia. Sua aplicação para determinação da estrutura 3D de biomoléculas, com destaque para as proteínas, deu origem à cristalografia de proteínas, caracterizada como um processo complexo que engloba uma variedade de estratégias e métodos tradicionais e modernos, integrando especialidades como a física, química, biologia, bioquímica e computação.

A cristalografia de proteínas determinou a criação de uma nova área do conhecimento, denominada biologia estrutural. A biologia estrutural encontra-se na interface entre a biologia molecular, a bioquímica e a biofísica, e tem como foco a investigação da estrutura de macromoléculas. A partir desta, busca-se elucidar a relação entre a estrutura e a função de uma determinada molécula. Por exemplo, a aplicação de métodos cristalográficos em macromoléculas biológicas permitiu o conhecimento da disposição dos átomos que constituem a estrutura 3D das moléculas de DNA, RNA e proteínas. Particularmente no caso desta última família de biomoléculas, além do entendimento do funcionamento dos organismos e das bases moleculares para a vida, as informações oriundas da cristalografia vêm sendo extremamente importantes no desenvolvimento de novos fármacos, como no caso de inibidores da protease do HIV e de moduladores de proteínas acopladas à proteína G (GPCR, *G protein-coupled receptor*).

Os estudos cristalográficos são componentes fundamentais para o desenvolvimento e a aplicação de métodos em bioinformática, incluindo a modelagem molecular e o planejamento de fármacos baseado na estrutura de receptores (SBDD, *structure-based drug design*). De fato, diversos métodos em bioinformática utilizam como pré-requisito o conhecimento 3D detalhado da macromolécula em estudo. Essa informação é geralmente adquirida a partir de estruturas depositadas em bases de dados públicos, onde podem ser acessadas livremente, dentre os quais se destaca o PDB (*Protein Data Bank*).

Embora a estrutura 3D de macromoléculas pode ser obtida através de diversos métodos experimentais, tais como a ressonância magnética nuclear (RMN, ver capítulo 12) e a criomicroscopia eletrônica, a cristalo-

grafia de raios-X ocupa papel de destaque. Isto pode ser evidenciado, por exemplo, no fato de que em janeiro de 2014 o PDB apresentava aproximadamente 97.000 estruturas de macromoléculas depositadas (incluídas proteínas, ácidos nucleicos, complexos macromoleculares e polissacarídeos), dentre as quais aproximadamente 90% tiveram sua estrutura 3D determinada pelo método de cristalografia de raios-X (Tabela 1-13).

Os métodos e estratégias cristalográficas para o estudo de macromoléculas evoluiram significativamente nos últimos anos. Devido aos rápidos avanços tecnológicos, as coletas de dados cristalográficos que eram realizadas exclusivamente em fontes caseiras (por exemplo, através de um ânodo rotatório) passaram a ser executada em fontes de alto brilho e intensidade, tais como laboratórios de luz síncrotron. Essa evolução tem como resultado direto um crescimento exponencial no número de estruturas de macromoléculas determinadas anualmente, conforme verificado pelo número de estruturas depositadas no PDB (Figura 1-13). Além disso, esse cenário tem contribuído para o desenvolvimento de duas abordagens distintas para o estudo de macromoléculas: *i*) tradicional e *ii*) larga escala.

A abordagem tradicional consiste em resolver estruturas de um pequeno conjunto de macromoléculas e seus complexos em um ambiente onde há ampla integração dos estudos cristalográficos com métodos bioquímicos, biofísicos e de química medicinal. Atualmente, projetos extremamente desafiadores em cristalografia têm como foco a determinação das estruturas de vírus, proteínas de membrana e complexos multimoleculares (por exemplo, envolvendo proteína-proteína, proteína–DNA e proteína–RNA).

Já a abordagem em larga escala consiste na elucidação do genoma estrutural através da determinação da estrutura 3D do maior número possível de proteínas constituintes de um determinado organismo. O desenvolvimento da cristalografia em larga escala (*high-throughput crystallography*) foi substancialmente beneficiado pelo surgimento de métodos automatizados para a cristalização e coleta de dados, bem como pelo desenvolvimento de fontes de luz de alto brilho e intensidade (por exemplo, síncrotrons de 3ª geração como o *European Synchrotron Radiation Facility* – ESRF, na França, o *Advanced Photon Source* – APS, nos EUA e o SPring-8, no Japão).

As estruturas 3D de proteínas determinadas por métodos cristalográficos são frequentemente o ponto de partida para a construção de modelos moleculares que visam elucidar a estrutura e função de proteínas homólogas (como visto no capítulo 7) ou o planejamento de novas moléculas bioativas (como visto no capítulo 9). Portanto, o co-

Tabela 1-13: Estruturas de macromoléculas depositadas no PDB (estatísticas de janeiro/2014).

Método experimental	Proteínas	Ác. nucleicos	Complexos proteína-DNA/RNA	Outras macromoléculas	Total
Cristalografia	79.922	1.497	4.162	4	85.585
RMN	8.990	1.065	197	7	10.259
Microscopia eletrônica	496	51	170	0	717
Híbridos	55	3	2	1	61
Outros	153	4	6	13	176
Total	89.616	2.620	4.537	25	96.768

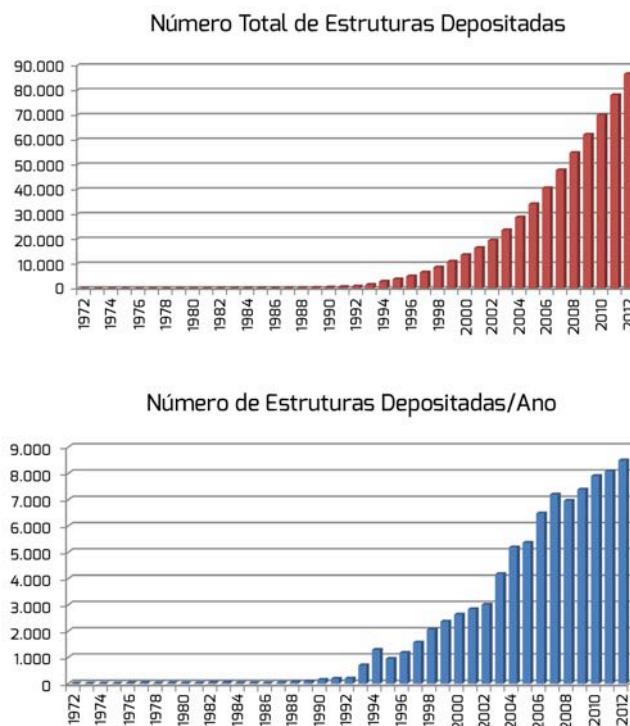



Figura 1-13: Número de estruturas de macromoléculas depositadas no PDB (dados 1972–2014, http://www.rcsb.org).

nhecimento dos fundamentos, vantagens e limitações da cristalografia de raios-X é fundamental para a seleção criteriosa de estruturas apropriadas para os estudos em bioinformática.

Adicionalmente, esse conhecimento nos permite uma melhor compreensão e avaliação dos modelos 3D de macromoléculas depositados nos bancos de dados. Desse modo, o presente capítulo busca oferecer uma descrição dos métodos cristalográficos para a determinação da estrutura 3D de proteínas, explorando seus princípios e fundamentos, com especial destaque para os critérios que devem ser utilizados para a obtenção de uma estrutura por cristalografia de raios-X, bem como para avaliação da qualidade do modelo estrutural construído.

13.2. Obtenção de proteínas

Uma das etapas fundamentais da biologia estrutural é a obtenção do alvo molecular em quantidade e pureza suficiente para os estudos cristalográficos (em torno de miligramas de proteína com teor de pureza maior que 95%).

Para contornar este desafio, os projetos pioneiros de cristalografia de macromoléculas (por exemplo, na cristalização da mioglobina em 1958, da hemoglobina em 1960, da lisozima em 1965 e da insulina em 1969) utilizaram proteínas extraídas de fonte natural (nos casos mencionados, músculo esquelético de baleia cachalote, sangue de cavalo, clara de ovo de galinha, pâncreas de porco, respectivamente). Entretanto, a utilização de fontes naturais para obtenção da macromolécula geralmente inclui algumas limitações, dentre as quais destacam-se:

i) baixa concentração: a pequena quantidade de proteína produzida na células, somada à distribuição diferenciada nos tecidos do organismo em estudo acarretam em baixa concentração de proteína para os estudos estruturais;

ii) isoformas e modificações pós-traducionais: a expressão de isoformas de uma proteína, aliada aos diferentes níveis de modificações pós-traducionais, aumentam a heterogeneidade da amostra e dificultam a separação dos componentes da solução. Essas características apresentam impacto significativo na obtenção de proteína com elevado teor de pureza e, consequentemente, na qualidade e formação dos cristais.

Apesar dessas limitações, algumas proteínas específicas continuam sendo obtidas a partir de fontes naturais, com destaque para anticorpos, proteínas de membrana e proteínas fúngicas envolvidas no processo de produção do bioetanol. Contudo, a vasta maioria das proteínas investigadas por métodos cristalográficos são provenientes de sistemas heterólogos (isto é, expressão realizada em organismo hospedeiro diferente do organismo alvo) baseados em estratégias de expressão que utilizam a tecnologia do DNA recombinante.

O avanço das técnicas de DNA recombinante e engenharia genética, com destaque para o desenvolvimento da reação em cadeia

da polimerase (PCR, *polymerase chain reaction*) permitiram o desenvolvimento de métodos de expressão heteróloga altamente eficientes para a produção de proteína pura e homogênea para os estudos estruturais. O emprego dessa tecnologia determinou mudanças significativas nos paradigmas da cristalografia de proteínas, permitindo que a investigação estrutural de proteínas, anteriormente baseada em baixíssima quantidades de proteína obtidas no organismo alvo ou dependentes do metabolismo celular, pudesse ser conduzida rotineiramente. Portanto, o domínio de técnicas e métodos bioquímicos e de biologia molecular tornaram-se componentes essenciais para a determinação estrutural de macromoléculas biológicas.

Nas próximas seções serão apresentados os métodos mais utilizados para produção de proteína em sistema de expressão heterólogo para os ensaios de cristalização. Contudo, é importante mencionar que, embora existam protocolos disponíveis para todas as etapas envolvidas (por exemplo, clonagem, expressão, purificação e cristalização), adaptações podem e devem ser realizadas para atender as particularidades da proteína em estudo.

A montagem de um sistema de expressão heteróloga necessita inicialmente do fragmento de DNA responsável pela codificação da proteína em estudo. De modo geral, a pesquisa minuciosa de informações da literatura indica dados relevantes para o desenvolvimento de protocolos otimizados de obtenção da proteína alvo. Nesse sentido, um protocolo de produção de uma proteína homóloga pode ser adaptado e utilizado como ponto de partida para o desenvolvimento de um novo método de obtenção da proteína de interesse. Na ausência desse tipo de informação qualificada, dados bioquímicos e moleculares como ambiente molecular da proteína *in vivo*, presença de parceiros fusionados, domínios estruturais, presença de regiões flexíveis e peptídeos de sinalização são extremamente úteis para o planejamento da nova construção genética.

Por exemplo, a descrição detalhada dos domínios constituintes de uma proteína é uma informação valiosa que contribui substancialmente para o desenvolvimento de um sistema de expressão heterólogo robusto. Domínios proteicos, tipicamente, apresentam capacidade de enovelamento independente, logo construções gênicas contendo somente um domínio podem ser estabelecidas.

Além disso, pode-se utilizar dados moleculares para truncar um domínio em posições específicas e, assim, remover alças flexíveis que dificultariam o processo de cristalização. Portanto, o planejamento da construção gênica deve ser realizado com base nos conhecimentos adquiridos sobre o alvo molecular em estudo e nos objetivos específicos que se deseja alcançar. Nesse contexto, é fortemente recomendada a utilização de ferramentas de bioinformática para auxiliar o planejamento de construções genéticas de alta eficiência.

Um exemplo de aplicação do conhecimento molecular no desenvolvimento de construções gênicas para estudos cristalográficos pode ser observado nos receptores nucleares. Estes receptores são proteínas multidomínios de grande interesse científico, pois exercerem função central no controle da expressão gênica. A complexa organização estrutural dos receptores nucleares, representada pelos seus diferentes domínios estruturais (Figura 2-13), exigiu uma análise detalhada para a obtenção de construções gênicas capazes de expressar de modo eficiente os diferentes segmentos. As construções planejadas expressaram com sucesso os domínios isolados dos receptores nucleares, tais como o domínio de complexação ao ligante do receptor RAR (PDB ID 3LBD) e o domínio isolado de ligação ao DNA do receptor GR (PDB ID 3FYL), bem como a estrutura integral do receptor PPARγ (PDB ID 3DZU) que, além de revelar a organização estrutural do receptor, confirmou a integridade e relevância dos domínios isolados.

As informações funcionais e estruturais, extremamente necessárias para elaboração de construções gênicas eficientes, podem ser usualmente obtidas através de métodos de

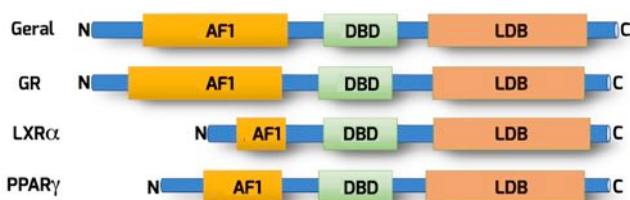



Figura 2-13: Distribuição representativa dos domínios de receptores nucleares GR (receptor de glicocorticoide, do inglês *glucocorticoid receptor*), LXRα (receptor hepático Xα, do inglês *liver Xα receptor*) e PPARγ (receptor γ ativado por proliferador de peroxissomo, do inglês *peroxisome proliferator-activated receptor γ*). N indica o domínio N-terminal, que contém a região com a função de ativação (AF, do inglês *activation function*), o domínio de ligação ao DNA (DBD, do inglês *DNA binding domain*) e o domínio de complexação ao ligante (LBD, do inglês *ligand binding domain*).

bioinformática. Por exemplo, há diversos métodos disponíveis para predição de propriedades moleculares importantes, como distribuição de estrutura 2ária, reconhecimento de domínios, presença de peptídeos de sinalização, hélices transmembranares, ligações dissulfeto intramoleculares, regiões flexíveis e desordenadas, dentre outras.

Construções gênicas

O planejamento e a montagem de construções gênicas para obtenção de proteínas envolvem diversos métodos de manipulação de DNA e sistemas de expressão. Dentre as diversas abordagens disponíveis para tal, duas estratégias de clonagem serão discutidas adiante: *i*) clonagem clássica em sistema de expressão bacteriano, e *ii*) clonagem em sistema independente de ligação – LIC (*ligation-independent cloning*). Adicionalmente, estes métodos vêm sendo facilitados pela disponibilidade cada vez maior de DNA sintético para aquisição diretamente de empresas especializadas.

A clonagem clássica inicia-se com o planejamento dos oligonucleotídeos iniciadores (*primers*) e da seleção do DNA molde. Os oligonucleotídeos iniciadores são utilizados para a amplificação por PCR do gene de interesse a partir do DNA molde. Geralmente, utiliza-se DNA genômico para organismos procarióticos e bibliotecas de DNA complementar (cDNA) para organismos eucarióticos (Figura 3-13).

O sucesso na amplificação do gene é verificado através de análise eletroforética em gel de agarose. Após purificação, procede-se com a ligação do fragmento amplificado em vetor de clonagem (por exemplo, TOPO® – Invitrogen). Vetores de clonagem apresentam alto número de cópias por célula e são utilizados para a transformação de bactérias específicas, tais como DH5α, Dh10B e XL1blue, as quais são empregadas para propagação do gene de interesse e fornecimento de DNA plasmidial. O fragmento de interesse é excisado do material obtido através da digestão com endonucleases de restrição. Essas enzimas reconhecem sequências de nucleotídeos específicas, inseridas no fragmento pelos oligonucleotídeos iniciadores, gerando terminais coesivos ou *stick ends*.

O fragmento isolado, obtido por separação eletroforética, é posteriormente ligado em vetor de expressão. A família de vetores e derivados do sistema pET® (Novagen) estão entre os mais utilizados para essa finalidade. Esses vetores são previamente tratados com as mesmas endonucleases para a criação de terminais complementares ao fragmento, o qual é ligado ao vetor com auxílio de uma DNA ligase. O plasmídeo elaborado é então introduzido em bactérias de propagação e, após confirmação da integridade da construção gênica, os plasmídeos são utilizados para a transformação de cepas bacterianas específicas para expressão proteica.

O método clássico é bastante robusto e amplamente empregado como alternativa atrativa na clonagem de genes. Contudo, inclui diversas etapas e detalhes que limitam sua aplicação em média e larga escala. Nesse sentido, tendo em vista a necessidade de aumentar a taxa de sucesso na obtenção de proteína expressa na forma solúvel, com alta pureza e em grande quantidade, novas estra-

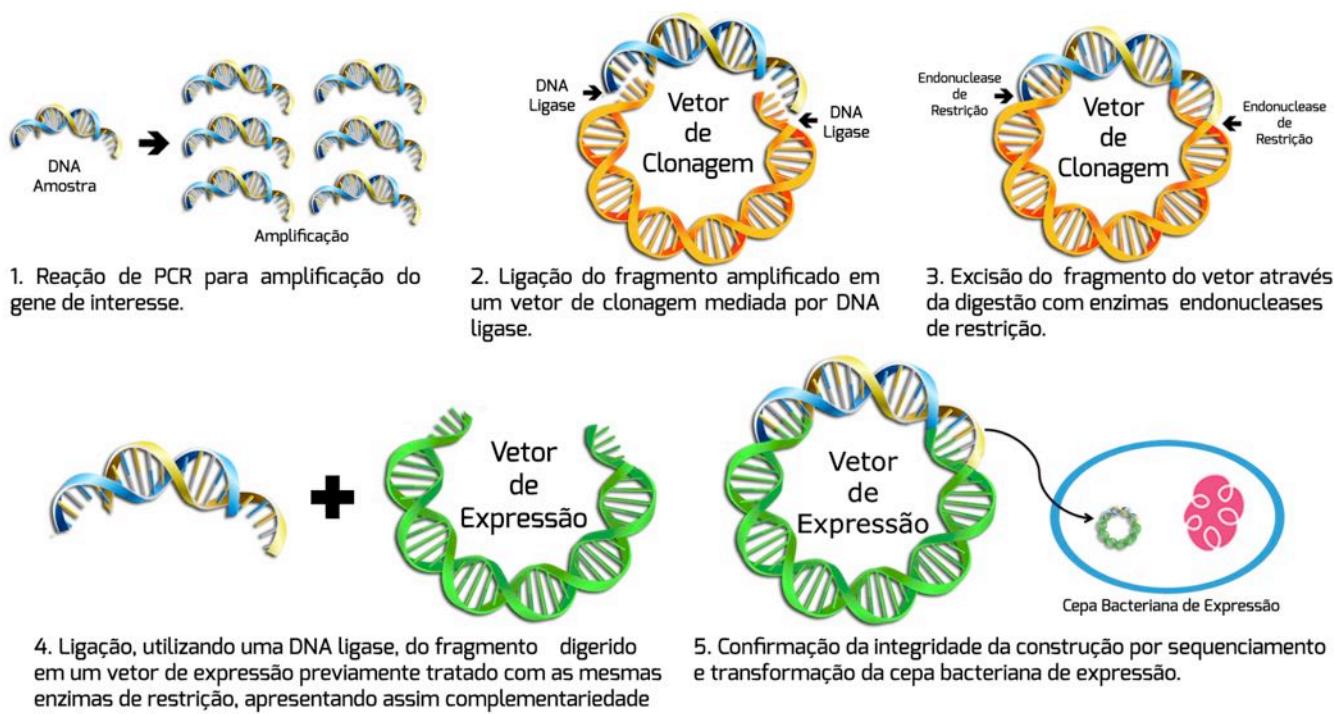



Figura 3-13: Esquema geral do método de clonagem clássica para expressão heteróloga de proteína.

tégias em biologia molecular, capazes de explorar diferentes possibilidades de expressão, foram desenvolvidas.

As construções gênicas planejadas passaram então a ser desenvolvidas em paralelo, aumentando-se as chances de sucesso na obtenção de proteína com as características adequadas para os estudos cristalográficos empregando o denominado sistema de clonagem independente de ligação (LIC) (Figura 4-13).

O sistema LIC diferencia-se do sistema clássico pela independência de uma etapa de ligação com DNA ligase. Adicionalmente, em algumas adaptações desse sistema pode-se evitar também o uso de endonucleases de restrição. Além disso, apresenta como vantagens: *i*) facilidade no planejamento do oligonucleotídeo iniciador, que inclui uma sequência específica do sistema para determinado conjunto de vetores, *ii*) disponibilidade de um número significativo de vetores preparados para este sistema, e *iii*) versatilidade na obtenção de construções gênicas variadas, não havendo a necessidade de etapas adicionais ou particularidades na utilização de um vetor determinado.

Em linhas gerais, após a amplificação e obtenção do fragmento de interesse através da reação de PCR com os oligonucleotídeos iniciadores específicos, trata-se o fragmento com a enzima T4 DNA polimerase na presença de um único tipo de nucleotídeo (por exemplo, dATP). A T4 DNA polimerase possui atividade exonuclease 3'–5' intrínseca, logo esta aplicação favorece a formação de extremidades salientes ou *overhangs*, complementares aos vetores utilizados. Em seguida, o fragmento é adicionado ao vetor escolhido, previamente tratado com T4 DNA polimerase e mantido em contato a temperatura ambiente. Por fim, essa mistura é utilizada na transformação da bactéria de propagação. Devido ao número de bases que são emparelhadas entre vetor e fragmento, através de suas saliências, não se faz necessária a utilização da DNA ligase, sendo a ligação covalente entre vetor e fragmento estabelecida pelo próprio sistema de reparo da bactéria transformada.

1.Reação de PCR para amplificação do gene de interesse.

2. Tratamento do fragmento amplificado com a enzima T4 DNA polimerase.

3. Ligação do fragmento e vetor tratado, independente de ligase.

4. Confirmação da integridade da construção por sequenciamento e transformação de cepa bacteriana de expressão.

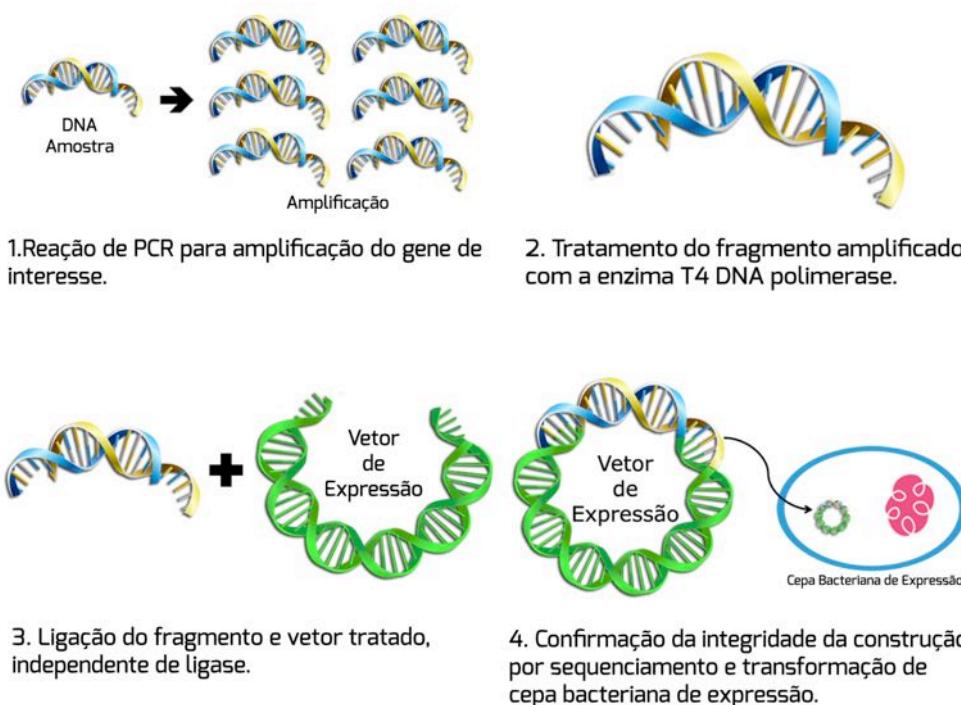



Figura 4-13: Esquema geral do método de clonagem independente de ligação (LIC) para expressão heteróloga de proteína.

13.3. Expressão

Atualmente, a expressão heteróloga é a fonte primária de produção de proteínas. Exemplos de organismos hospedeiros que "emprestam" sua maquinaria celular para a expressão proteica incluem bactérias, protozoários, fungos, células de insetos e de mamíferos e sistema de expressão independente de célula hospedeira (*cell-free expression*), também conhecido como expressão *in vitro*.

Em um experimento padrão de expressão heteróloga de proteína as células hospedeiras são cultivadas até atingirem uma biomassa crítica, medida pela densidade óptica (DO) da cultura. A partir desse momento inicia-se o procedimento de indução da expressão da proteína de interesse. Nos vetores bacterianos, um dos mecanismos para controle de indução é o operon *lac*, de forma que a presença de lactose ou derivados (como a alolactose) favorece a indução da expressão da proteína através da ligação da alolactose ao repressor do operon. Análogos otimizados da alolactose foram desenvolvido e, dentre eles, o derivado mais utilizado é o isopropil-1-tiol-β-D-galactopiranosídeo (IPTG). O IPTG se liga ao repressor *lac* e induz a superexpressão da proteína de interesse. Como a bactéria não é capaz de metabolizá-lo, a concentração do agente indutor permanece constante, favorecendo a manutenção dos níveis de expressão.

Parâmetros como meio de cultura, aeração, densidade óptica antes da indução, concentração de agente indutor, temperatura e tempo de expressão afetam significativamente a produção de proteína solúvel. Dentre eles, a temperatura e a concentração do agente indutor estão entre os parâmetros de maior impacto sobre a expressão e, portanto, devem ser cuidadosamente avaliados. Tipicamente, experimentos conduzidos em temperaturas mais baixas (menores que 37°C) determinam uma redução na taxa de expressão. Contudo, favorecem a obtenção de proteína enovelada corretamente.

Paralelamente, diferentes concentrações do agente indutor devem ser testadas para a identificação das condições ideais que determinam um nível de expressão adequado para os estudos cristalográficos. Entretanto, frequentemente, a proteína de interesse não é obtida na forma solúvel, seja pelo enovelamento incorreto ou pelo acúmulo em corpos de inclusão. Nesses casos, pode-se recuperar a proteína da fração insolú-

vel através de técnicas de solubilização dos corpos de inclusão, como através do emprego de detergentes, e de re-enovelamento (*refolding*).

Por outro lado, se o enovelamento não foi atingido com sucesso ou a proteína expressa é não funcional devido à ausência de modificações pós-traducionais, uma alternativa é a expressão em células eucarióticas. Para esses casos são recomendados sistemas de expressão em células de fungo, protozoário, mamífero ou inseto.

A escolha do sistema de expressão (vetor + organismo de expressão) depende de vários fatores. Por exemplo, em relação ao vetor de expressão, dependente do organismo de expressão, há diversas opções disponíveis com estruturas moleculares similares, mas que diferem em relação ao mecanismo de regulação, sítios de restrição, antibiótico de resistência, presença de proteínas acessórias e facilitadores de purificação.

Em relação à escolha do organismo de expressão, um dos aspectos mais importantes a ser considerado consiste na necessidade de modificações pós-traducionais, isto é, modificações na estrutura proteica após síntese como enovelamento mediado por chaperonas, formação de pontes dissulfeto, glicosilação e etc. Por exemplo, o sistema bacteriano (procariótico) não é capaz de glicosilar proteínas de eucariotos. Portanto, caso seja necessária a realização desta ou modificações pós-traducionais não realizadas por bactérias deve-se optar por sistemas mais adequados para essa finalidade.

Devemos observar que a ausência de modificações pós-traducionais pode determinar a produção de uma proteína não funcional ou, até mesmo, enovelada incorretamente. Por outro lado, estratégias de cristalização podem explorar características como a incapacidade do sistema bacteriano de realizar glicosilações como as vistas em eucariotos. Nesse sentido, a ausência de modificações pós-traducionais pode ser benéfica para o processo de cristalização, uma vez que alterações desse tipo aumentam a heterogeneidade intrínseca da proteína em solução, tendo impacto direto no processo de cristalização.

Sistema de expressão em bactérias

O sistema de expressão mais utilizado é o bacteriano, sendo a *Escherichia coli* o organismo de primeira escolha para expressão de proteína para estudos cristalográficos. A *E. coli* é responsável pela produção de mais de 85% das proteínas depositadas no PDB (dados jan/2014), fato relacionado às características do organismo, tais como: *i*) crescimento rápido; *ii*) baixa virulência; *iii*) facilidade de manipulação; *iv*) elevada produção de proteínas recombinantes.

Atualmente, existe uma variedade significativa de cepas modificadas e otimizadas para expressão bem sucedida de proteínas recombinantes, com destaque para aquelas derivadas da cepa BL21, Rosetta™ (Novagen®), Origami™ (Novagen®), B834 (Novagen®) e cepas que apresentam o plasmídeo pLysS.

A cepa Rosetta™ possibilita rendimentos elevados na produção de proteínas eucarióticas que apresentam códons raros. Essa característica está relacionada à presença do plasmídeo pRARE, que suplementa a bactéria com RNAs transportadores (RNAt) para esses códons.

A cepa Origami™ é indicada para aumentar o rendimento de proteína enovelada e funcional dependente da formação de ligações dissulfeto. Para tanto, possui mutantes das proteínas tiorredoxina redutase e glutationa redutase que favorecem a formação dessas ligações no citoplasma.

A cepa B834 e similares, auxotróficas para a produção de metionina, são úteis para a produção de proteínas contendo o aminoácido modificado selenometionina, apresentando-se como alternativa atrativa e relevante para a determinação estrutural de proteínas como, por exemplo, na obtenção experimental de fases utilizando o sinal anômalo do átomo Se.

Por fim, as cepas que contém o plasmídeo pLysS são adequadas para a produção de proteínas tóxicas para a bactéria. A presença do plasmídeo determina que os níveis de expressão basais sejam reduzidos ao máximo, evitando-se assim danos celulares.

Sistema de expressão em fungos

As células fúngicas têm sido ampla-

mente empregadas como um bem sucedido sistema de expressão alternativo para proteínas de interesse cristalográfico. Entre as cepas mais populares destacam-se as leveduras *Saccharomyces cerevisiae* e *Pichia pastoris*, além dos fungos filamentosos *Aspergillus niger* e *Trichoderma reesei*.

As principais características da utilização das células fúngicas para expressão consistem em: *i*) baixo custo para o cultivo; *ii*) elevada densidade celular, embora necessite de um tempo maior para obtenção da densidade adequada quando comparado às bactérias; *iii*) rendimento satisfatório, alcançando desde mg/L até g/L de cultivo; *iv*) possibilidade de modificações pós-traducionais; *v*) introdução de marcadores para secreção da proteína no meio de cultura.

Em geral, a cepa selecionada direciona a montagem da construção gênica. Sendo assim o vetor, o marcador molecular de secreção da proteína de interesse e o padrão de modificações pós-traducionais são específicos para a cepa utilizada. Além disso, os procedimentos e infraestrutura para o emprego desse sistema são mais sofisticados e demandam maior tempo, havendo necessidade de avaliar os transformantes para encontrar uma cepa com níveis de expressão elevados.

Sistema de expressão em células de mamíferos

A produção de proteína recombinante em células de mamíferos é realizada com sucesso em alguns casos, produzindo-se proteínas funcionais especialmente quando os alvos são de origem humana. As linhagens celulares comumente empregadas para expressão de proteína são as células embriônicas de rim humano 293 (HEK 293, *human embryonic kidney* 293), células do ovário de hamsters (CHO, *chinese hamster ovary*) e COS (célula tipo fibroblastos derivadas de rim de macaco).

A principal vantagem desse sistema de expressão consiste na obtenção de proteínas complexas enoveladas corretamente, por exemplo, como no caso do segmento extracelular da integrina αVβ3, PDB ID 1JV2. Dentre as limitações, contudo, pode-se mencionar: *i*) custo elevado de produção, devido às particularidades do cultivo desse tipo celular e o baixo rendimento obtido; *ii*) incapacidade de produção de proteínas tóxicas para o hospedeiro; *iii*) dificuldade de adaptação a sistemas de triagem em larga escala (HT, *high-throughput*).

Sistema de expressão em células de insetos

Uma alternativa para produção de proteínas em células de mamíferos é a utilização de células de insetos, capazes de realizar modificações pós-traducionais similares àquelas promovidas por células de mamíferos.

A principal linhagem celular utilizada é a *Spodoptera frugiperda*, sendo a expressão mediada pela infecção das células por um baculovírus que funciona como o vetor de expressão. Dentre as vantagens desse sistema, em relação às células de mamíferos, citam-se: *i*) maior rendimento na produção de proteína recombinante; *ii*) pode ser adaptado para ensaios HTS; *iii*) possibilidade de trabalhar com linhagens adequadas à cultura em suspensão, permitindo o uso de biorreatores.

13.4. Purificação

A pureza da amostra é um dos principais fatores que influenciam o processo de cristalização de macromoléculas. Nesse sentido, é fortemente recomendável que a proteína em estudo apresente o maior teor de pureza possível, sendo essa característica dependente de procedimentos de purificação robustos e eficazes. Estes, por sua vez, dependem da estratégia de clonagem e sistema de expressão da proteína.

A primeira etapa do processo de purificação é a lise da célula de expressão. O processo de lise celular é bastante crítico pois, dependendo das condições no qual é realizado (tais como o método de lise, agente tamponante, pH, presença de cofatores, detergentes e temperatura) a proteína pode ser degrada-

da ou acumular-se na fração insolúvel. Assim, faz-se necessário avaliar criteriosamente as melhores condições de lise.

Frequentemente, a etapa seguinte consiste na precipitação fracionada das proteínas na mistura proveniente da lise celular. Esse procedimento é realizado através da adição de um sal, como o sulfato de amônio, ou de um solvente orgânico, como o etanol. Com os avanços das técnicas e métodos de expressão recombinante, vetores de expressão modernos permitem a inclusão de facilitadores da purificação. Nesse sistema, as proteínas são expressas com marcadores (*tags*) que possibilitam o emprego de métodos cromatográficos (particularmente cromatografias de afinidade) para a captura seletiva da proteína de interesse.

O tipo de método cromatográfico a ser empregado depende do marcador vinculado ao vetor do sistema de expressão. Esses marcadores variam desde oliogopeptídeos, como uma cauda de hexahistidina (6xHis), até proteínas fusionadas de elevada massa molecular, como a glutationa-S-transferase (GST). A cromatografia de afinidade por íon metálicos imobilizados é comumente utilizada para purificação de proteínas expressas com cauda de hexahistidina.

Após a etapa de cromatografia de afinidade deve-se decidir sobre a manutenção ou remoção do marcador. Não há evidências claras sobre o impacto do marcador para o processo de cristalização, contudo, geralmente remove-se os marcadores de elevada massa molecular e avalia-se a influência dos de pequena massa molecular.

A remoção do marcador ou clivagem é realizada pelo tratamento da amostra com enzimas proteolíticas, como trombina, fator Xa, enteroquinase, TEV protease e SUMO protease. A seleção da enzima é determinada pela estratégia de clonagem e vetor utilizado, uma vez que este contém sequências de reconhecimento específicas para determinadas proteases.

Nesse momento, uma segunda etapa de cromatografia de afinidade deve ser utilizada para separar a proteína de interesse dos marcadores e das proteínas não digeridas pela protease. Subsequentemente, uma etapa de cromatografia de exclusão por tamanho, também conhecida por gel filtração, é necessária para a purificação final da amostra.

O método de gel filtração permite ainda a avaliação da homogeneidade da amostra em relação aos estados oligoméricos existentes em solução, o que pode ter implicações importante na compreensão da biologia estrutural da proteína em estudo. Além disso, pode-se empregar essa técnica para realizar a troca da solução tamponante para uma mais adequada para os ensaios de cristalização.

É importante mencionar que, além da cromatografia de afinidade e de gel filtração, outros métodos cromatográficos são frequentemente empregados para aumentar o teor de pureza da proteína em estudo, tais como a cromatografia de troca iônica e a cromatografia de interação hidrofóbica. Essas técnicas são aplicadas à amostra proteica nos casos em que a pureza obtida não tenha atingido os níveis necessários para os estudos cristalográficos.

O teor de pureza recomendado para cristalografia de proteínas é superior a 95%. Contudo, faz-se necessário esclarecer que a cristalização é, em si, um método de purificação, de forma que não há regra absoluta sobre a pureza da amostra. Comumente, avalia-se a pureza da proteína através de análise eletroforética desnaturante em gel de poliacrilamida (SDS-PAGE), cujo resultado ideal é a presença de uma banda única correspondente à proteína de interesse na forma pura (Figura 5-13). Métodos alternativos como análises eletroforéticas não desnaturantes e ensaios de espalhamento dinâmico de luz (DLS, *dynamic light scaterring*) são frequentemente empregados para assegurar o teor de pureza e homogeneidade da solução em estudo.

13.5. Cristalização

A obtenção de cristais adequados para os experimentos de difração de raios-X é fundamental para a determinação da estrutura

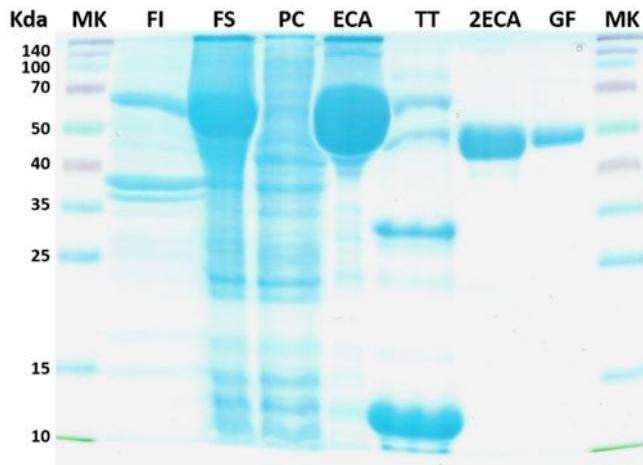



Figura 5-13: Gel representativo de análise eletroforética desnaturante em SDS-PAGE para a enolase de *Plasmodium falciparum*. Da esquerda para direita estão apresentados o marcador de massa molecular (MK), a fração insolúvel (FI), a fração solúvel (FS), a passagem livre pela coluna de afinidade (PC), a eluição da coluna de afinidade (ECA), o tratamento com TEV protease (TT), a eluição da segunda coluna de afinidade (2ECA) e a eluição da gel filtração (GF).

tridimensional de macromoléculas. O fenômeno de cristalização ocorre quando a molécula em estudo precipita de modo lento e ordenado, formando cristais (Figura 6-13). O processo ocorre em condições controladas, incluindo uma solução supersaturada da proteína de interesse, agentes precipitantes, condições controladas de temperatura, força iônica e em pequenos intervalos de variação de pH.

Os cristais são caracterizados por arranjos periódicos constituídos de unidades formadoras, que podem variar desde uma única molécula até grandes complexos macromoleculares, tais como ribossomos ou ainda um capsídeo viral.

As interações químicas entre as moléculas que constituem as unidades formadoras de cristais proteicos são de baixa energia, tais como interações dipolo-dipolo, interações por ligação de hidrogênio, interações eletrostáticas e interações de van der Waals. Como resultado dessa rede de interações de baixa energia e alto conteúdo de solvente (~50%), cristais de proteínas mostram-se extremamente frágeis quando comparados a cristais de sais inorgânicos.

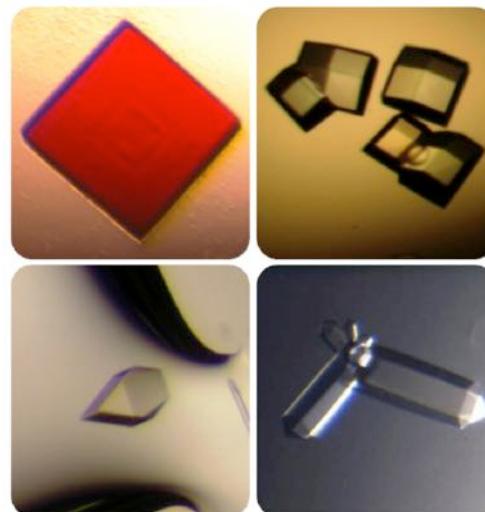



Figura 6-13: Exemplos de cristais de proteínas.

O tamanho dos cristais de proteína é bastante variável, com dimensões entre 1 e 500 μm. Adicionalmente, suas características macroscópicas são, na maioria das vezes, consequência da ordem (ou simetria do grupo espacial) no qual as moléculas se empacotaram no retículo cristalino.

As propriedades da proteína, como distribuição de cargas na superfície, presença de regiões flexíveis e distribuição de conformações têm impacto significativo no fenômeno de cristalização. Esse processo ocorre a partir de uma solução supersaturada de proteína, sendo a velocidade com que se atinge esse estado essencial para a formação de cristais, microcristais ou precipitado amorfo.

A cristalização de macromoléculas biológicas é uma técnica baseada na tentativa e erro por se tratar de um processo complexo e multiparamétrico. Parâmetros de caráter físico (como temperatura, pressão, superfície da molécula e tempo) e químico (como pH, agente precipitante, forca iônica, grau de supersaturação, pureza da amostra, estado de agregação, ponto isoelétrico e presença/ausência de estabilidade) interferem diretamente na formação de um cristal, de maneira que os diversos métodos utilizados exploram esse espaço multiparamétrico com o objetivo de examinar os efeitos de combinações dessas

variáveis. Esses métodos são geralmente aplicáveis à maioria das proteínas, DNAs, RNAs e complexos multimoleculares.

Dentre os parâmetros que podem apresentar impacto direto no processo de cristalização merece destaque a temperatura, capaz de alterar a curva de solubilidade da proteína e a cinética de equilíbrio e nucleação. As temperaturas amplamente empregadas para cristalização de proteínas são de 18 °C e 4 °C embora, quando possível, recomenda-se avaliar a influência de temperaturas alternativas.

No processo de cristalização, a vasta maioria das interações entre as moléculas de proteínas ocorrem na superfície das mesmas. Portanto, a presença ou ausência de algumas características podem ser fundamentais para obtenção de um cristal, destacando-se a presença de regiões desordenadas ou muito flexíveis e distribuição dos resíduos superficiais que contribuem para a carga total e entropia do sistema. A distribuição de algumas propriedades, calculadas a partir da sequência de aminoácidos do alvo proteico, como número de aminoácidos, ponto isoelétrico, tamanho da maior região desordenada, estabilidade, presença de domínios *coiled coil*, entre outras, tem sido empregada na avaliação do potencial de cristalização ou cristalizabilidade. Ferramentas computacionais, como o XtalPred, avaliam essas propriedades e as comparam com aquelas disponíveis em banco de dados de proteínas cristalizadas para predizer a capacidade da proteína de interesse de cristalizar.

Independentemente da origem e das particularidades da macromolécula em estudo, alguns parâmetros importantes favorecem a produção de cristais adequados aos estudos de difração de raios-X, com destaque para: *i*) a quantidade de proteína, necessária para garantir amostra suficiente durante os experimentos, e *ii*) a pureza da amostra. Embora existam casos de cristalização a partir de extratos brutos (como é o caso da lisozima, da ferritina e da mioglobina), a probabilidade de sucesso nos experimentos de cristalização aumenta significativamente com o emprego de proteína com elevado teor de pureza.

A solução de proteína inicialmente empregada em ensaios de cristalização apresenta concentração abaixo do limite de solubilidade, ou seja, constitui uma solução insaturada. Logo para que a cristalização ocorra é necessário que essa solução se torne supersaturada (Figura 7-13). Nesse sentido, deve-se aumentar a concentração da solução de proteína através da remoção do solvente e da inclusão de agentes precipitantes, capazes de reduzir a solubilidade da proteína. O sistema então evoluirá para um estado mais concentrado, que ultrapassará o limite de solubilidade e constituirá uma solução supersaturada.

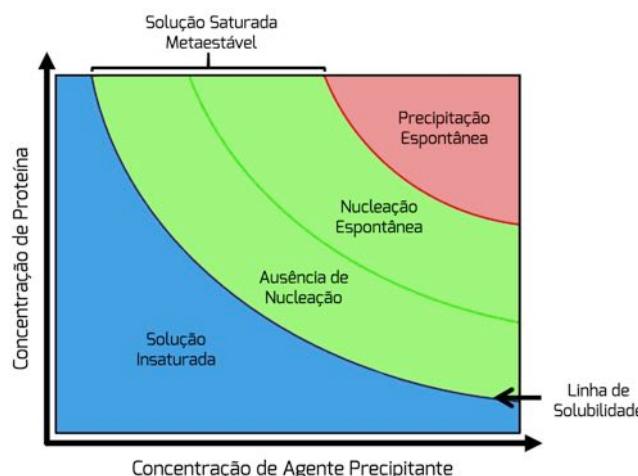



Figura 7-13: Diagrama de fase mediado por agente precipitante e concentração proteica para a cristalização.

A análise do diagrama de fase representado na Figura 7-13 revela três regiões distintas:

i) região azul, caracterizada pela presença de solução insaturada (proteína solúvel). Nessa região não há formação e crescimento de cristais;

ii) região verde, caracterizada pela solução saturada metaestável, subdivida nas sub-regiões *iia* e *iib*;

iia) abaixo da linha central verde não haverá formação e crescimento de cristais devido à ausência de núcleos cristalinos;

iib) acima da linha verde a formação de cristais torna-se favorável, pois ocorre o fenômeno de nucleação de maneira espontânea. Nessa região a barreira energética é vencida, permitindo que o sistema reti-

re proteína da solução e forme os núcleos cristalinos. Este processo é acompanhado pela diminuição da concentração de proteína em solução, e o sistema evoluirá para o equilíbrio que favorece o crescimento dos cristais a partir dos núcleos formados;

iii) região vermelha, caracterizada pela presença de solução hipersaturada. Nessa região ocorre precipitação espontânea da proteína de forma desordenada.

As condições favoráveis para o processo de nucleação e crescimento de cristais devem ser avaliadas cuidadosamente. Nesse contexto, podem-se identificar condições favoráveis para o crescimento do cristal que, contudo, não são favoráveis para a nucleação. Da mesma forma, pode-se obter condições favoráveis para a nucleação intensa da proteína que, por sua vez, impedem o crescimento dos cristais. Existem diversas técnicas para contornar os problemas específicos de cada caso, buscando-se a obtenção de cristais adequados para os estudos cristalográficos.

Processo físico-químico

A cristalização pode ser descrita como um processo físico-químico envolvendo os seguintes componentes energéticos:

$$\Delta G_{crist} = \Delta H_{crist} - T(\Delta S_{proteína} + \Delta S_{solvente})$$

O termo entálpico (ΔH_{crist}) apresenta contribuições modestas ao processo de cristalização, uma vez que é proveniente de um pequeno número de interações moleculares de baixa intensidade, estabelecidas entre as macromoléculas para a formação do cristal.

Paralelamente, esse processo determina a perda de liberdade de translação e rotação das macromoléculas quando comparadas às suas formas livres em solução. Perde-se ainda a flexibilidade de algumas alças devido ao empacotamento estabelecido sendo, portanto, um processo entropicamente desfavorável ($\Delta S_{proteína} > 0$).

Por outro lado, a cristalização da macromolécula libera uma quantidade significativa de moléculas de águas previamente ordenadas ao redor de resíduos hidrofóbicos e polares, o que promove um ganho entrópico considerável ($\Delta S_{solvente} < 0$) que torna o processo de cristalização espontâneo ($\Delta G_{crist} < 0$).

A compreensão dos componentes energéticos é de fundamental importância para o favorecimento do processo de cristalização. Atualmente, altera-se a capacidade de cristalização de proteínas através de mutações específicas de resíduos localizados na superfície da macromolécula de forma a interferir nestes componentes, favorecendo a cristalização. Exemplos relevantes dessa estratégia incluem modificações de resíduos de aminoácidos com termo entrópico elevado, especialmente, resíduos de lisinas e ácidos glutâmicos. Estes resíduos possuem cadeias laterais longas e, por sua disposição preferencial pela superfície proteica, normalmente caracterizam-se por elevada entropia conformacional. Desta maneira, a troca por resíduos com menor entropia associada, como exemplo resíduos de alanina, minimizam a perda de entropia durante o empacotamento, favorecendo ainda mais o processo de cristalização ($\Delta G_{crist} << 0$).

O planejamento de mutações com objetivo de aumentar o potencial de cristalização de um alvo macromolecular é auxiliado pela disponibilidade de servidores gratuitos na internet. Um exemplo importante é o SERp da Universidade da Califórnia (UCLA). Esse servidor emprega o método de redução da entropia de superfície (SER, *surface entropy reduction*) que, em linhas gerais, realiza a previsão de estrutura 2^{ária} a partir da sequência de aminoácidos e, com base nesse resultado, estabelece o perfil entrópico da proteína, sugerindo resíduos cuja mutação poderia beneficiar o processo de cristalização.

Métodos de cristalização

Uma vez obtida a proteína de interesse com teor de pureza adequado, tem-se diversas alternativas disponíveis para a cristalização. Em comum, estes métodos envolvem a mistura da solução pura de proteína com soluções de cristalização, contendo agentes precipitantes variados.

Em seguida, mantém-se a mistura em um sistema fechado e isolado para estabelecimento do equilíbrio e consequente cristalização. A seleção da estratégia de cristalização depende de fatores como o objetivo de aplicação (por exemplo, a triagem inicial de condições ou a otimização de cristais) e características do ensaio (como a facilidade de resgate dos cristais da gota de cristalização, o número de experimentos e a possibilidade de automação, dentre outros).

O método de difusão de vapor baseia-se no equilíbrio entre duas soluções através da fase de vapor em sistema fechado. A solução menos concentrada perde seu solvente volátil até que os potenciais químicos das duas soluções se igualem. Para se controlar a concentração final da solução de proteína, realiza-se o experimento de difusão de vapor com um volume pequeno de proteína contra um volume grande de solução precipitante. Assim, uma gota de solução da proteína a ser cristalizada é adicionada à solução tampão contendo agentes precipitantes e aditivos (por exemplo, em uma diluição 1:1). Em seguida, essa gota é equilibrada contra um reservatório contendo uma solução de agentes precipitantes a uma concentração maior que a da gota com proteína. O método de difusão de vapor pode ser conduzido de duas maneiras principais: a gota suspensa (*hanging drop*) e a gota assentada (*sitting drop*) (Figura 8-13).

No método gota suspensa, a gota contendo a proteína de interesse é colocada sobre uma lamínula de vidro siliconizada e, posteriormente, vedada com o auxílio de graxa especial na parte superior do poço, como aquele em uma placa de 24 poços, de forma que a gota fique interna ao reservatório (Figura 8-13).

Entre as vantagens dessa metodologia destaca-se a facilidade e versatilidade de aplicação. Entre as limitações encontra-se o custo elevado das lamínulas, a impossibilidade de automação e a dificuldade de montagem das gotas quando um dos agentes precipitantes promove perda da tensão superficial.

No método gota assentada, a solução contendo a proteína é colocada sobre um suporte fixado no centro do poço, o qual é posteriormente vedado com o auxílio de fita adesiva apropriada (Figura 8-13).

Entre as principais vantagens desse método destaca-se a capacidade para desenvolvimento de experimentos automatizados e miniaturizados, com a utilização de placas de 96, 384 e 1536 poços, empregando com gotas de até 50 nL. Entre suas limitações tem-se o tempo de espera entre a montagem de uma gota e a etapa de vedação da placa, que deve ser suficientemente rápido para impedir que a gota evapore totalmente, e a possibilidade de alguns cristais ficarem aderidos à superfície da placa.

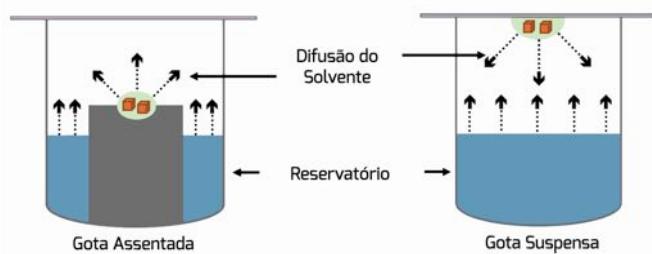



Figura 8-13: Métodos de cristalização que utilizam a técnica de difusão de vapor.

A escolha do método está associada ao propósito do ensaio. Assim, experimentos de triagem de condições de cristalização são tipicamente conduzidos com o emprego do método da gota assentada, enquanto para etapas de reprodução de cristais e otimização de condições utiliza-se o método da gota suspensa.

Adaptações e estratégias diferenciadas são frequentemente empregadas nesses métodos, buscando modificar os estados iniciais e finais do sistema e a cinética de equilíbrio. Por exemplo, podem ser empregadas proporções distintas de solução de cristalização e solução proteica, como 1:2, 2:1 e 1:3, além da utilização de óleos permeáveis e impermeáveis sobre a solução do reservatório.

Métodos alternativos de cristalização de proteínas incluem o *batch*, a microdiálise e a interfase livre de difusão (Figura 9-13).

O método *batch* emprega concentrações de solução de proteína e agentes precipitantes adequadas para gerar uma nova solução proteica supersaturada. A solução resultante é então coberta por óleo imper-

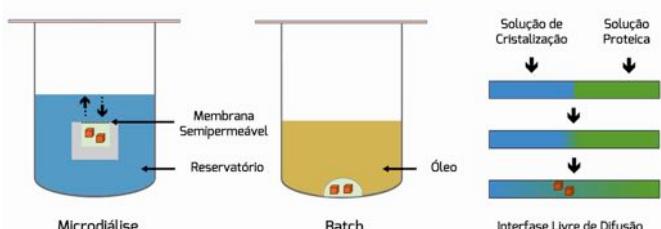



Figura 9-13: Exemplos de métodos de cristalização alternativos empregados em cristalografia de proteína.

meável, que dificulta a difusão de vapor e, dessa forma, isola o sistema para que se atinja o equilíbrio. Consequentemente, é favorecida a cristalização da macromolécula (Figura 9-13). Variantes dessa técnica utilizam óleos permeáveis, como silicones, que determinam novas condições de equilíbrio para a formação de cristais de boa qualidade.

A microdiálise permite a troca do solvente e do agente precipitante presente na solução proteica com a solução do reservatório através de uma membrana semipermeável, favorecendo a redução ou aumento das concentrações e, consequentemente, a cristalização.

Na interfase livre de difusão a solução de proteína e a solução de cristalização são acondicionadas em capilares que permitem o contato das soluções em apenas uma pequena superfície (interface de contato), de forma que o equilíbrio é atingido após a difusão lenta de uma solução na outra. Nesse experimento, avalia-se o perfil de solubilidade da proteína em gradiente de concentração para identificação da condição mais favorável para a cristalização.

A automatização das etapas de montagem e observação dos cristais tem favorecido significativamente os experimentos de cristalização, propiciando:

i) ganho de agilidade e precisão na montagem dos cristais, particularmente importantes em trabalhos com proteínas sensíveis e instáveis e na reprodutibilidade dos cristais;

ii) miniaturização;

iii) redução no custo e consequente possibilidade de aumento no número de ensaios realizados para o mesmo alvo;

iv) viabilização de estudos de cristalização para proteínas cuja expressão seja bastante reduzida ao permitir a manipulação dos pequenos volumes envolvidos.

Reagentes para cristalização

As soluções de cristalização contêm reagentes que podem ser agrupados em classes distintas: agentes tamponantes (responsáveis por manter o pH adequado da solução de cristalização), aditivos (facilitam e/ou otimizam o processo de cristalização) e precipitantes (reduzem a solubilidade da proteína).

O agente tamponante é fundamental no processo de cristalização por manter constante o pH da solução e, consequentemente, estabilizar a distribuição de cargas dos resíduos na superfície da proteína. Além disso, o agente tamponante pode alterar a solubilidade da proteína favorecendo o processo de cristalização quanto empregados em concentração adequada.

Os aditivos são compostos capazes de permitir, facilitar ou aperfeiçoar o processo de cristalização como, por exemplo, cloreto de magnésio, L-prolina, ATP e NAD. Esses compostos apresentam propriedades distintas, que favorecem o processo de cristalização. Por exemplo, detergentes estabilizam a estrutura e impedem a agregação de proteína, enquanto ligantes e íons metálicos podem promover contatos intermoleculares adicionais ou ainda alterar a polaridade do meio. Diante da impossibilidade de prever o efeito de determinado aditivo sobre a cristalização, deve-se avaliar a influência desses compostos através de triagem sistemática. Para tanto, há disponíveis kits comerciais já preparados para aplicação.

Os agentes precipitantes podem ser divididos em duas classes: sais inorgânicos e compostos orgânicos. A utilização de sais como agentes precipitantes está relacionada a dois fenômenos conhecidos como *salting-in* e *salting-out*. O primeiro favorece o aumento da solubilidade da proteína através do acréscimo de pequenas quantidades de sal, enquanto o segundo favorece a diminuição da solubilidade da proteína por acréscimos de quantidades elevadas de sal. Sais como o sulfato de amônio, cloreto de sódio e citrato de sódio estão entre os amplamente empregados como agentes precipitantes.

Na classe dos precipitantes orgânicos destacam-se os polímeros de poliálcoois, com ênfase para o polietilenoglicol (PEG) e polietilenoglicol monoetil éter (PEG–MME), que apresentam comprimentos de cadeias variáveis, variando de 200 a 20.000 Da de massa molecular média. Os representantes mais

populares dessa classe são os PEGs 3.350, 4.000 e 8.000. O mecanismo de redução de solubilidade por estes compostos é atribuído à competição dos substituintes hidroxilas com os resíduos da proteína pelas moléculas de água disponíveis.

Adicionalmente, alguns álcoois de pequena massa molecular têm sido empregados com sucesso como agentes precipitantes. Estes compostos são capazes de reduzir a concentração de proteína pela alteração da polaridade da solução de cristalização. Exemplos de destaque dessa categoria incluem o isopropanol, 2-metil-2,4-pentanodiol (MPD), 1,6-hexanodiol e glicerol.

Estratégias para cristalização de proteínas

Atualmente, as etapas iniciais de triagem para identificação de condições de cristalização promissoras empregam soluções de cristalização isoladas ou reunidas de acordo com as características físico-químicas. Essas soluções são produzidas e comercializadas por empresas especializadas, tais como Hampton Research, Molecular Dimensions, Qiagen e Jena Biosciences.

Dentre os formatos e estratégias disponíveis destaca-se a triagem em rede (*grid screen*), capaz de fornecer informações importantes de modo rápido, sendo por isso amplamente aplicada em triagens iniciais. Nesse experimento, avaliam-se sistematicamente dois fatores em paralelo como, por exemplo, variações simultâneas de pH/PEG, pH/cloreto de sódio e pH/sulfato de amônio, dentre outras combinações.

Uma estratégia alternativa para identificação de condições promissoras para a cristalização consiste na utilização de soluções fatoriais. Nessa abordagem, busca-se balancear a ocorrência de algumas características principais e suas combinações durante o processo de amostragem através do planejamento fatorial. A utilização de fatoriais incompletos reduz a quantidade de parâmetros avaliados e, com isso, o número de experimentos realizados. Essa alternativa encontra aplicação quando a disponibilidade de amostra restringe o número de ensaios que podem ser conduzidos.

Devido às características do processo automatizado de montagem dos experimentos de cristalização, a estratégia mais empregada em triagens iniciais é a matriz esparsa, que apresenta aspectos semelhantes ao fatorial incompleto. Para a elaboração dessa estratégia, um estudo estatístico que incluiu mais de 500 proteínas, 480 condições de cristalização e mais de 500.000 experimentos foi conduzido pelo centro de genômica estrutural *Joint Center for Structural Genomics* (JCSG – San Diego, Califórnia, EUA). Esse estudo resultou na seleção de 384 condições com maior probabilidade de sucesso para a cristalização de macromoléculas.

Para a realização dos ensaios de cristalização há necessidade de solução de proteína com a máxima pureza disponível e concentração adequada. A concentração média utilizada para determinação das estruturas de proteínas depositadas no PDB é de 14 mg/mL. No entanto, há exemplos de estruturas cristalizadas entre 2 e 100 mg/mL. Como regra geral, emprega-se a concentração de 10 mg/mL em ensaios iniciais de cristalização.

Após a montagem dos experimentos, as placas de cristalização devem ser acondicionadas em ambiente adequado, com baixa vibração e temperatura controlada, para que o sistema evolua em direção à condição de equilíbrio.

Tradicionalmente, a observação das gotas é realizada através de análise visual com o auxílio de uma lupa. Contudo, equipamentos modernos e programas de reconhecimento de padrões têm sido desenvolvidos e aplicados na inspeção e aquisição de imagens, onde fotos de cada uma das gotas do experimento de cristalização são obtidas e analisadas automaticamente. Como regra geral, observa-se o experimento imediatamente após sua montagem (t = 0), seguida de mais 10 observações ao longo do experimento, com intervalos menores no início e mais prolongados ao final.

As observações devem ser registradas adequadamente para avaliação e identificação das condições mais promissoras para cristalização. Os kits comerciais fornecem tabelas próprias com sistemas de pontuação para facilitar a interpretação e análise dos resultados. Adicionalmente, programas têm sido utilizados como

ferramentas eficientes para avaliação dos dados e proposição de novos experimentos.

O objetivo dos experimentos de cristalização é a obtenção de cristais adequados para os ensaios de difração de raios-X. No entanto, os resultados observados podem ser bastante variados, incluindo-se:

i) cristais bem formados, com arestas e faces definidas (Figura 10A-13);
ii) cristais com crescimento em duas dimensões, denominados de placas (Figura 10B-13);
iii) cristais com crescimento em apenas uma dimensão, denominados de agulhas (Figura 10C-13);
iv) precipitados leves e intensos (Figuras 10D-13 e 10E-13, respectivamente);
v) separações de fase (Figura 10F-13);
vi) aglomerados de agulhas (Figura 10G-13);
vii) microcristais (Figura 10H-13).

Com exceção de alguns casos nos quais os cristais obtidos na etapa de triagem podem ser considerados adequados para os experimentos de difração de raios-X, a obtenção de uma condição promissora é seguida por etapas de otimização. Embora o número de parâmetros a serem investigados nessa etapa seja elevado, costuma-se explorar a concentração dos reagentes iniciais (incluindo a concentração de proteína), a proporção entre a solução de proteína e a solução de cristalização, o agente tamponante e o pH da solução, a presença de aditivos e detergentes e a cinética de equilíbrio, entre outros. Essa investigação se estende até a identificação de condições otimizadas de cristalização, capazes de fornecer cristais apropriados e de boa qualidade para os experimentos de difração de raios-X.

13.6. Coleta de dados

Uma vez que cristais adequados são produzidos, eles podem ser testados quanto à sua capacidade de difração de raios-X e, em seguida, serem empregados na coleta de dados cristalográficos.

O uso da difração de raios-X na obtenção de informação sobre a estrutura de moléculas baseia-se na propriedade do padrão de difração da distribuição eletrônica dos átomos em um objeto poder ser aproximado pela transformada de Fourier do mesmo. Por outro lado, a transformada inversa de Fourier do padrão de difração é a distribuição eletrônica dos átomos do cristal de proteína.

O fenômeno de difração depende da interação entre a radiação eletromagnética com a matéria do objeto e da dispersão dessa radiação ao incidir sobre este. Embora existam outros métodos de dispersão disponíveis, como a dispersão de nêutrons dos núcleos, eles constituem atualmente uma fração muito pequena dos experimentos de difração.

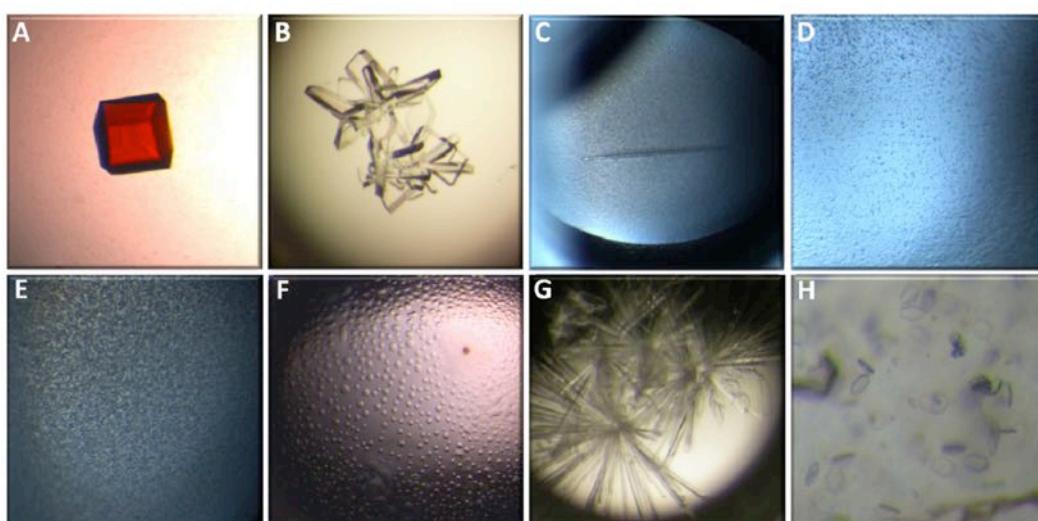



Figura 10-13: Resultados possíveis em experimentos de cristalização. A) cristal bem formado, B) placas, C) agulhas, D) precipitado leve, E) precipitado intenso, F) separação de fase, G) aglomerados de agulhas e H) microcristais.

Em relação às proteínas ou outras moléculas orgânicas, os raios-X são a radiação eletromagnética de escolha para os estudos estruturais. O comprimento de onda típico dos raios-X é de 0,15 nm (1,5 Å), ou seja, da mesma ordem do comprimento de uma ligação covalente entre átomos. Consequentemente, torna-se possível detectar tais distâncias, utilizando-se a difração de raios-X.

Em princípio, um único objeto já é capaz de difratar raios-X. Assim, uma única molécula seria suficiente para a realização de experimentos de difração de raios-X. No entanto, a utilização de uma única molécula como fonte espalhadora resulta em feixes de radiação dispersos de baixíssima intensidade, cuja detecção é praticamente impossível pelos métodos disponíveis.

Para solucionar essa limitação, utiliza-se uma quantidade significativa de moléculas (~10^{15} moléculas) organizadas num padrão regular tridimensional. Este grande número de moléculas atua como amplificador do sinal, capaz de gerar feixes de radiação mensuráveis de alta intensidade. Por conseguinte, estruturas cristalinas são as mais adequadas para obtenção de dados de alta resolução em experimentos de difração de raios-X.

Padrão de difração

O padrão de difração de proteínas é tridimensional e reflete tanto a simetria dos arranjos cristalinos quanto a organização da proteína na célula unitária, isto é, a unidade de repetição que constitui o cristal). Esses arranjos são definidos em termos de grupos espaciais e de unidades assimétricas.

A unidade assimétrica é a menor unidade a partir da qual uma célula unitária pode ser construída. Além disso, a unidade assimétrica representa o número mínimo de estruturas independentemente determinadas em um cristal. Por exemplo, uma unidade assimétrica pode conter desde apenas um representante da proteína em estudo até 12 ou mais representantes. Frequentemente, esses arranjos tornam possível a determinação do estado oligomérico da proteína, especialmente em casos nos quais as subunidades não são idênticas (Figura 11-13).

Para a determinação das coordenadas espaciais dos átomos da proteína, responsáveis pela difração do feixe de raios-X, faz-se necessário identificar cada uma das reflexões no padrão de difração (Figura 12-13). Devido ao caráter tridimensional do padrão de difração, as distâncias entre as reflexões medidas, em um detector, localizam-se próximas ou distantes do centro do padrão. Portanto, a partir de um ponto de origem (o centro da imagem) valores crescentes são atribuídos para todas as reflexões no padrão de difração. Esses valores, denominados índices de Miller, indicam reflexões próximas do centro da imagem (ou seja, valores menores de índices de Miller) e reflexões localizadas nas regiões periféricas do padrão de difração (ou seja, valores maiores índices de Miller).

Os ângulos que os feixes difratados fazem com relação ao feixe incidente no cristal determinam o nível de informação obtido em um experimento de difração de raios-X. Assim, quanto maior o ângulo do feixe difratado

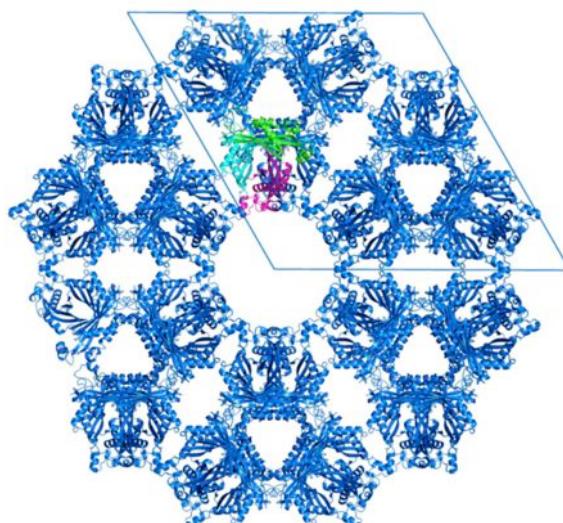



Figura 11-13: Exemplo de empacotamento cristalino, célula unitária (paralelogramo) e unidade assimétrica (destacada nas cores ciano, magenta e verde). Empacotamento de várias moléculas da proteína 6-piruvoil-tetrahidrobiopterina-sintase humana (PTPS). Dados processados e refinados por JRCM e gentilmente cedidos pelo *Structural Genomics Consortium*, Oxford, UK.

maiores serão os valores dos índices de Miller para as reflexões observadas, e por conseguinte, maior será a resolução dos dados coletados (Figura 12A-13).

Informações moleculares a alta resolução produzem mapas de densidade eletrônica bem definidos, que auxiliam substancialmente a determinação precisa da posição dos átomos que constituem o cristal (Figura 12B-13). Portanto, os detalhes e qualidade do modelo 3D da macromolécula são diretamente proporcionais à resolução dos dados coletados nos estudos cristalográficos.

Fundamentalmente, as características do padrão de difração, isto é, as intensidades das reflexões e a resolução do conjunto de dados, determinam a qualidade do mapa de densidade eletrônica. Nesse sentido, parâmetros quantitativos são empregados para avaliação da qualidade do conjunto de dados, dentre os quais destacam-se a intensidade das reflexões (I), os danos causados pela radiação (Rdano), a sobreposição das reflexões (O), o fator R (Rmerge) e a completeza (C) (Tabela 2-13).

Figura 12-13: (A) Padrão representativo obtido em um experimento de difração de raios-X de uma estrutura de altíssima resolução (1,1 Å). Os anéis vermelhos indicam as camadas de resolução para as reflexões. As reflexões se tornam menos intensas quanto maior a resolução. (B) Resolução *versus* densidade eletrônica. Mapa de densidade eletrônica para o mesmo resíduo de triptofano calculado em 4 diferentes resoluções (PDB ID 3T7L). Dados de difração gentilmente cedidos pelo *Structural Genomics Consortium*, Oxford, UK.

Intensidade (I)

As intensidades das reflexões têm impacto direto na qualidade dos dados cristalográficos. A intensidade das reflexões depende de diversos fatores, tais como o tamanho e a qualidade do cristal, o tempo de exposição ao feixe de raios-X e a intensidade do feixe de raios-X.

A relação entre a intensidade da reflexão e o plano de fundo (*background*) é dada pela razão sinal-ruído $I/\sigma(I)$. Uma vez que as proteínas estão sujeitas a alterações causadas pela interação com raios-X, causadas por radicais livres, durante a coleta de dados cristalográficos deve-se ponderar a relação entre o tempo de exposição do cristal e a intensidade do feixe de modo que se obtenham intensidades mensuráveis e de boa qualidade, sem afetar a estrutura da proteína em estudo.

Tais limites de resolução dos dados de difração são frequentemente definidos pelo critério $I/\sigma(I)$. Em geral, utiliza-se dados que apresentam valores de $I/\sigma(I)$ maiores que 2, isto é, a intensidade medida para as reflexões é duas vezes maior que o ruído observado.

Rdano (R)

O valor de Rdano indica a extensão do impacto das colisões do tipo inelásticas e elásticas provenientes do feixe de fótons incidentes na amostra cristalina. Devido à alta intensidade desses fótons a amostra sofrerá processos irreversiveis e será "danificada.

Os danos causados pela radiação constituem um importante fator para a qualidade dos dados cristalográficos. Com o objetivo de amenizar tais danos, geralmente é empregada uma estratégia de coleta de dados a temperaturas "criogênicas" (100 K), obtidas com o auxílio de nitrogênio líquido.

A aplicação dessa estratégia para coleta de dados cristalográficos exige um pré-tratamento do cristal. Cristais de proteína contém uma quantidade significativa de água, logo seu resfriamento acarreta na formação de gelo que, por sua vez, é extremamente prejudicial para o cristal e, consequentemente, para o experimento de difração.

Por este motivo os cristais são usualmente pré-tratados com agentes crioprotetores, tais como PEG

Tabela 2-13: Dados cristalográficos representativos de uma coleta de dados de difração de raios-X (PDB ID 3ZRS).

Dados Cristalográficos	
Grupo espacial	$P\,4\,2_1\,2$
Dimensões da célula (Å)	a = b = 106,24 c = 89,80 $\alpha = \beta = \gamma = 90º$
Resolução (Å)	106,24 – 3,05 (3,21-3,05)*
Rmerge	0,262 (0,945)*
$\langle I \rangle / \langle \sigma(I) \rangle$	5,5 (2,0)*
Completeza (%)	99,9 (99,8)*
Multiplicidade	6,8 (6,9)*

*Os números entre parênteses referem-se à mais alta camada de resolução.

ou glicerol, seguidos de resfriamento rápido (*flash cooling*). Este procedimento evita a formação de cristais de gelo, mantendo assim a integridade e qualidade dos cristais de proteína.

Sobreposição [O]

Além da intensidade da reflexão, a capacidade para discernir reflexões individuais também é essencial. A separação das reflexões em um padrão de difração depende, principalmente, do tamanho da célula unitária. Nesse sentido, quanto maior as dimensões da célula unitária (parâmetros a, b e c da Tabela 2-13) mais próximas estarão as reflexões no padrão de difração e consequentemente, maior será a probabilidade de ocorrer sobreposição.

Esta sobreposição de reflexões acarreta em uma maior imprecisão na determinação da intensidade de cada reflexão. Além disso, outros fatores como a desordem interna no cristal (mosaicidade), proveniente do empacotamento cristalino ou de danos mecânicos (como aqueles causados durante o resfriamento rápido) podem ocasionar alargamento significativo das reflexões no padrão de difração produzindo sobreposição.

Rmerge [Rm]

Uma vez que o padrão de difração contém os elementos de simetria do cristal, a maioria das reflexões é observada mais de uma vez. Dessa maneira, a reprodutibilidade dessas medidas é uma característica utilizada como parâmetro de precisão.

Estatisticamente, quanto maior a frequência com que uma reflexão é medida, e quanto mais similares elas são entre si, melhor será o conjunto de dados cristalográfico. A redundância desses dados é indicada em termos de uma média geral, enquanto a reprodutibilidade das medidas é avaliada por um fator residual denominado Rmerge (ou Rsym, quando se leva em conta a simetria das reflexões).

O valor de Rmerge é obtido através do cálculo da média da intensidade de um grupo de reflexões dividido pela média do desvio padrão para esse mesmo grupo de reflexoes:

$$R_{merge} = \frac{\sum_h \sum_i |I_i - \langle I \rangle|}{\sum_h \sum_i I_i}$$

É importante mencionar que o fator Rmerge é dependente da resolução, logo deve ser informado para todo o conjunto assim como para as camadas de mais altas de resolução (Tabela 2-13). Um conjunto de dados de boa qualidade caracteriza-se por um valor de Rmerge global menor que 15% e, na camada de maior resolução, o valor de Rmerge dever ser menor que 100%.

Completeza [C]

A completeza dos dados é um fator extremamente importante na determinação da qualidade do conjunto. A completeza é determinada pela razão entre o número esperado de reflexões para o grupo espacial e o tamanho da célula unitária. Uma vez que a capacidade para medir reflexões diminui em função da resolução, a completeza dos dados será menor nas camadas de maior resolução. Portanto, esse parâmetro deve ser informado tanto para todo o conjunto de dados quanto para a camada mais alta de resolução (Tabela 2-13).

Um conjunto de dados cristalográficos ideal é formado por camadas de baixa e alta resolução determinadas com relação sinal-ruído ($I/\sigma(I)$) global maior que 10 e maior que 2 para a camada de maior resolução, reflexões bem separadas, valor de Rmerge global

menor que 100% e completeza maior que 95% (em geral, é aceitável que a completeza seja baixa somente nas camadas de maior resolução).

A relação entre esses parâmetros determina a qualidade final do mapa de densidade eletrônica. Portanto, quanto maior a qualidade dos dados cristalográficos, maior será a probabilidade de se obter um mapa de densidade eletrônica bem definido e interpretável. No entanto, é importante mencionar que a análise isolada desses parâmetros não deve ser utilizada como um substituto para o julgamento da veracidade do modelo estrutural.

Os valores mencionados para os principais parâmetros cristalográficos devem ser utilizados como indicativos da qualidade do conjunto de dados coletados. A vasta maioria dos modelos estruturais depositados no PDB foi construído a partir de conjuntos de dados de excelente qualidade. Contudo, há também exemplos de modelos incorretos, provenientes de conjuntos de dados de qualidade simplesmente aceitável. Em geral, esses modelos são resultado da interpretação inadequada dos mapas de densidade eletrônica, construídos a partir de conjunto de dados de menor resolução. Portanto, quanto maior a resolução dos dados, menor a probabilidade de erros no modelo estrutural da proteína em estudo.

Faseamento

A radiação eletromagnética pode ser descrita pela equação de ondas, que é definida em termos de amplitude, comprimento de onda e fase. Em um experimento de difração de raios-X, os dois primeiros parâmetros são medidos diretamente, ou seja, a amplitude da onda é proporcional à intensidade do feixe difratado (a amplitude é igual à raiz quadrada da intensidade medida para uma reflexão) e o comprimento de onda (λ) é definido pelo comprimento de onda dos raios-X utilizados. As fontes caseiras com ânodo rotatório de Cu apresentam λ = 1,54178 Å, enquanto fontes de luz síncrotrons apresentam λ = 0,8–2,5 Å.

A determinação da fase nos estudos cristalográficos é um processo complexo, conhecido como "problema das fases". É uma etapa fundamental e de grande impacto para a obtenção de mapas de densidade eletrônica bem definidos e, por conseguinte, para a construção de modelos estruturais de qualidade. De fato, um mapa de densidade eletrônica calculado a partir das amplitudes de uma estrutura correta, mas com fases incorretas, seria impossível de se interpretar. Por outro lado, um mapa de densidade eletrônica calculado a partir de amplitudes de estruturas aleatórias, mas com fases corretas, seria interpretável.

A fase corresponde ao tempo relativo à chegada da crista de uma onda específica a um ponto de referência. Ondas de mesmo comprimento e fases idênticas terão seus picos e vales em comum, somando-se em harmonia. Ondas com fases opostas tendem a anular umas as outras, total ou parcialmente, dependendo de suas amplitudes.

Assim, ao somarmos todas as ondas difratadas (a síntese de Fourier) para se resolver uma estrutura de proteína, torna-se necessário determinar as amplitudes e fases para cada uma das ondas espalhadas, ou seja, para cada reflexão.

Experimentalmente, a amplitude da onda difrata é facilmente medida utilizando-se detectores modernos, tais como placas de imagem, *couple charged devive* (CCD) e *pixel apparatus for the SLS* (PILATUS). Em um experimento de difração, as intensidades e posições das ondas difratadas são medidas, mas as fases são perdidas. Isto ocorre porque os raios-X deslocam-se na velocidade da luz e, dessa maneira, o tempo relativo de chegada de todas as ondas espalhadas provenientes do cristal ao detector parece ser o mesmo. Portanto, as fases deverão ser determinadas através de métodos alternativos.

O método mais comum de faseamento, especialmente para o desenvolvimento de novos compostos bioativos, é o de substituição molecular. O método baseia-se em dois fatores: 1) na disponibilidade das coordenadas atômicas da estrutura da proteína de interes-

se ou a de uma proteína homóloga, e 2) na semelhança do padrão de difração da proteína de interesse com o padrão de difração da proteína homóloga.

Na substituição molecular, medem-se as amplitudes de difração do cristal da proteína de interesse e "substituem-se" as fases desconhecidas pelas fases já calculadas a partir de uma estrutura previamente determinada. A questão crucial que determina o sucesso deste método é o nível de semelhança entre as duas proteínas. Por exemplo, ao determinarmos a estrutura de um complexo ligante-proteína, esperamos que a interação do ligante com o sítio de ligação induza apenas alterações locais na estrutura do sítio, sem consequências maiores para a estrutura geral da proteína.

Nesses estudos, utilizam-se as amplitudes coletadas do cristal contendo o complexo proteína-ligante combinadas com as fases da proteína sem o ligante, previamente determinada. Esse método resulta em um mapa de densidade eletrônica para a proteína e para o ligante suficientemente adequado, permitindo a identificação do modo de interação do candidato a fármaco no sítio de ligação do alvo macromolecular (Figura 13-13).

Além da substituição molecular, é importante mencionar que existem outros métodos para a determinação das fases, tais como a substituição isomórfica e o espalhamento anômalo. Esses métodos são geralmente empregados nos casos em que a substituição molecular não é bem sucedida ou quando não há uma estrutura relacionada.

Mapa de densidade eletrônica

O mapa de densidade eletrônica é o resultado final de um experimento de difração de raios-X. Por definição, o mapa de densidade eletrônica é a solução da síntese de Fourier com as amplitudes das difrações medidas e as fases experimentalmente determinadas ou calculadas para cada reflexão. A partir deste mapa, procede-se para a etapa de interpretação e construção do modelo estrutural.

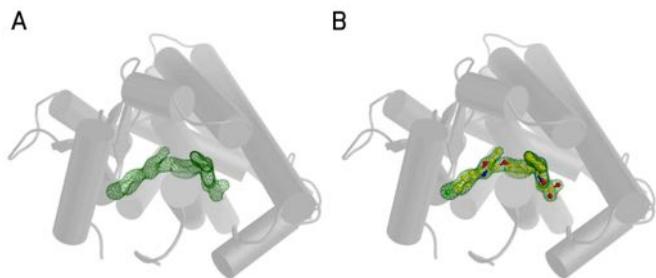



Figura 13-13: Estrutura do receptor PPARα complexado ao ativador NKS (PDB ID 3KDU). (A) Mapa de densidade eletrônica (malha verde), indicando o modo de interação do ativador NKS. (B) Complexo NKS-PPARα, no qual o ligante (esfera e bastões amarelos) encontra-se modelado de acordo com o mapa de densidade eletrônica.

Há disponíveis diversas operações que podem ser aplicadas aos dados cristalográficos com o objetivo de melhorar os mapas de densidade eletrônica. Uma estratégia frequentemente empregada é o achatamento do solvente (*solvent flattening*), que acentua as fronteiras entre o solvente e a molécula, tendo como resultado final a otimização do mapa de densidade eletrônica.

Adicionalmente, quando há mais de uma molécula na unidade assimétrica, a promediação (isto é, interpolação) das suas densidades eletrônicas pode aumentar a relação sinal-ruído, melhorando a qualidade do mapa final.

A interpretação do mapa de densidade eletrônica é subjetiva, demandando habilidade e experiência para que o modelo construído explique da melhor maneira possível os dados cristalográficos. Um dos fatores que interferem nesta interpretação é a resolução, que indica o nível de detalhamento com o qual a proteína foi determinada.

Níveis de resolução distintos determinam diferentes tipos de informação (Tabela 3-13 e Figura 12-13). O valor médio de resolução dos modelos estruturais depositados no PDB é 2 ± 1 Å, sendo que aproximadamente 40% das macromoléculas depositadas tem resolução entre 1,5–2,0 Å (dados de dezembro de 2012). Portanto, o mapa de densidade eletrô-

Tabela 3-13: Relação entre a informação estrutural e a resolução de um dado conjunto de dados cristalográficos.

Resolução	Informação estrutural
5,0	Topologia da molécula e elementos de estrutura secundária
3,5	Curso geral da cadeia polipeptídica (traço de $C\alpha$)
3,0	Cadeias laterais de alguns aminoácidos são interpretáveis
2,5	Cadeias laterais de todos aminoácidos são interpretáveis
1,5	Átomos individuais são reconhecíveis
1,0	Tipos de átomos são identificáveis

nica nessa faixa de resolução é rico em informação estrutural e facilmente interpretável e, por conseguinte, o modelo final construído tende a apresentar boa qualidade.

Diversos fatores contribuem para a facilidade de interpretação de um mapa de densidade eletrônica. Uma vez que a densidade eletrônica é uma média das posições atômicas ao longo de todas as células unitárias que formam o cristal, um mapa de densidade eletrônica nítido depende do perfeito alinhamento entre todas as moléculas.

Um mapa de densidade eletrônica inequívoca corresponde a apenas uma molécula, resíduo, modelo peptídico ou ligante que poderá ser modelado nessa densidade eletrônica. No entanto, se a densidade eletrônica não é bem definida, mas difusa, ou se houver moléculas em diferentes orientações, a interpretação se torna desafiadora.

Por exemplo, a cadeia lateral de um resíduo de aminoácido em um peptídeo pode adotar mais de uma conformação. Se o número de conformações for pequeno, como dois rotâmeros, essas conformações são modeladas com ocupações fracionadas (isto é, 50% para cada uma) (Figura 14-13). Se o número de conformações for significativo, com um número de rotâmeros > 3, a densidade eletrônica para esses rotâmeros não será distinguível, e aparecerá como ruído no mapa.

Um fenômeno semelhante é observado quando um ligante interage com apenas algumas moléculas de proteína no cristal. Nesse caso, o mapa de densidade eletrônica será fraco para esse ligante devido à ocupação parcial, sendo portanto de difícil interpretação e modelagem. A ocupação dos átomos no cristal é indicada em termos fracionários, que variam entre 0 e 1.

A incerteza associada à posição média dos átomos constituintes do cristal é indicada por um termo denominado fator B ou fator de temperatura. Quanto maior o deslocamento espacial dos átomos no cristal, maior será o fator B. Esse termo é dependente da resolução do conjunto de dados, apresentando valores médios para átomos em uma proteína no intervalo de 20–30 $Å^2$.

A ocupação e o fator B estão relacionados entre si, bem como a resolução do conjunto de dados. Geralmente, em complexos ligante-proteína é comum a verificação de fatores B significativamente maiores para os átomos do ligante em relação aos átomos da proteína, fenômeno este que pode indicar uma ocupação parcial para a molécula do ligante.

Mapas de densidades eletrônicas podem ser exibidos de diversas maneiras. A representação mais comum para a interpretação empregam os coeficientes $F_O - F_C$ e $2F_O - F_C$. O mapa $F_O - F_C$ indica a diferença entre a den-

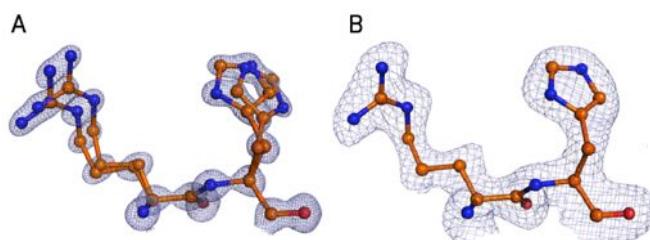



Figura 14-13: Exemplo de dupla conformação do mesmo segmento de uma proteína em diferentes resoluções (PDB ID 2VB1). (A) Dupla conformação em uma estrutura refinada na ultraresolução de 0,65 Å. Nota-se que as densidades eletrônicas adotam um formato de elipsoides, típico em casos de ultraresolução. As duplas conformações para os resíduos de arginina e histidina foram modeladas com precisão. (B) Mesma estrutura resolvida a 2,0 Å de resolução. Entretanto, apesar da boa qualidade dos dados não foi possível modelar as duas conformações adotadas por esses resíduos.

sidade eletrônica observada (F_O) e a calculada a partir de um modelo (F_C). Esse mapa, conhecido como "mapa diferença", evidencia regiões no modelo que necessitam de átomos, isto é, a diferença na densidade eletrônica é positiva, e regiões no modelo que apresentam excesso de átomos, ou seja, a diferença na densidade eletrônica é negativa.

O mapa $2F_O - F_C$ apresenta a densidade eletrônica com ênfase na diferença entre a densidade eletrônica observada ($2F_O$) e a calculada a partir de um modelo (F_C) (Figura 15-13). Durante o processo de refinamento do modelo cristalográfico, deve-se avaliar e interpretar de forma integrada os mapas $2F_O - F_C$, que privilegiam os fatores de estrutura observados, e o mapa diferença $F_O - F_C$, que indica regiões com excesso ou ausência de densidade eletrônica.

Figura 15-13: Mapa de densidade eletrônica $2F_O - F_C$ (malha azul) e $F_O - F_C$ (malha verde para densidade positiva e malha vermelha para densidade negativa). (A) O resíduo de arginina foi modelado em uma conformação que não condiz com os dados experimentais (densidades positivas e negativas no mapa $F_O - F_C$). (B) Rotâmero modelado corretamente para o mesmo resíduo de arginina. Nota-se que as densidades no mapa diferença desapareceram, indicando o acerto no posicionamento do rotâmero de arginina. Além disso, uma nova molécula de água (cruz vermelha) também foi corretamente modelada após seleção do rotâmero correto para o resíduo.

13.7. Refinamento, validação e usos

Os modelos estruturais construídos baseados em dados cristalográficos devem ser, idealmente, modelos precisos. Para tanto, diversos métodos de refinamento são empregados.

Uma estratégia comum de refinamento aplicada a modelos cristalográficos é o alinhamento correto entre o modelo estrutural e a densidade eletrônica. Esse processo é realizado de forma sistemática e supervisionado por ciclos interativos de refinamento no espaço real e no espaço recíproco. Para avaliação do protocolo de refinamento, consideram-se os parâmetros denominados Rfator e Rlivre (*Rfree*). Os ciclos de refinamento são conduzidos continuadamente até que ocorra convergência dos dados, ou seja, o processo de refinamento estende-se até o momento em que não se observam variações significativas nos valores de Rfator e Rlivre.

Com o objetivo de auxiliar o refinamento, restrições estereoquímicas são aplicadas para orientar o grau de liberdade conformacional dos átomos durante as tentativas de modelá-los na densidade eletrônica da proteína. Desse modo, garante-se a não violação das geometrias permitidas para os diferentes grupos químicos, bem como impede-se que a molécula adote conformações de alta energia. Essas restrições são baseadas no conhecimento estrutural de pequenas moléculas elucidadas a alta resolução e utilizadas como subestruturas representativas da macromolécula (Figura 16-13).

O sucesso no processo de refinamento é indicado pelo parâmetro Rfator, que consiste na medida de concordância entre o modelo construído e os dados experimentais. O valor de Rfator determina a diferença entre as amplitudes das reflexões calculadas derivadas a partir do modelo e os valores experimentais obtidos a partir do experimento difração de raios-X. Portanto, o valor de Rfator indica a qualidade do ajuste do modelo a densidade eletrônica, bem como a qualidade dos dados cristalográficos.

Para proteínas, os valores de Rfator observados encontram-se no intervalo de 15 a 20% para conjuntos de dados entre 1,8 e 2,5 Å de resolução (Figura 17-13). Esses números sugerem que de 75 a 80% dos dados de espalhamento, provenientes do cristal da proteína,

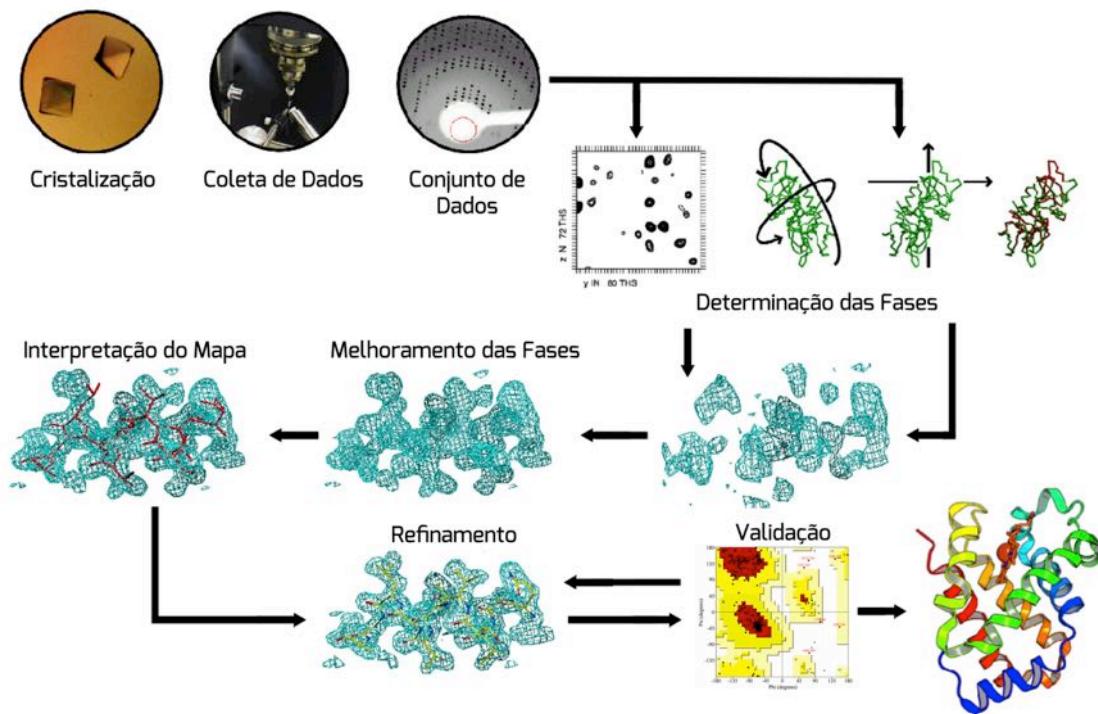



Figura 16-13: Visão geral das etapas envolvidas na determinação de uma estrutura de proteína por métodos cristalográficos.

podem ser representados ou explicados pelo modelo estrutural.

É importante mencionar que um modelo estrutural de boa qualidade pode apresentar pequenas falhas, provenientes de erros durante a aquisição dos dados cristalográficos, da incapacidade de se modelar regiões desordenadas na estrutura, de diferentes conformações e de regiões flexíveis, principalmente regiões de alças.

Devido à grande influência das fases calculadas (F_C) sobre as amplitudes das reflexões (F_O) na determinação da densidade eletrônica final, o valor de Rfator pode ser manipulado e levar ao sobreajuste do modelo estrutural.

Visando-se manter a precisão e a veracidade do modelo estrutural, uma estratégia comumente utilizada consiste no cálculo do Rfator a partir de dados que não foram utilizados no processo de refinamento e, portanto, não foram influenciados pelas fases calculadas, o que pode ser chamado de validação externa ou Rlivre.

O Rlivre é calculado a partir de 5 a 10% das reflexões, selecionadas de modo aleatório e excluídas do processo de refinamento. Devido à natureza incompleta dos dados utilizados para o cálculo do Rlivre, este é frequentemente maior do que o valor do Rfator em cerca de 3–5%, no caso de estruturas bem refinadas. Nas etapas iniciais de refinamento, esse número pode ser maior que 10%.

Uma vez que as moléculas de proteína são formas irregulares, durante o processo de formação dos cristais espaços e canais entre as cadeias polipeptídicas são preenchidos com solvente e outros compostos provenientes da solução de cristalização, incluindo-se água, íons e agente crioprotetor, dentre outros.

O componente mais importante do solvente são as moléculas de água ligadas à proteína, encontradas em localizações discretas e, geralmente, na superfície da macromolécula. As moléculas de água são modeladas de acordo com um procedimento que envolve a identificação de características específicas das densidades eletrônicas que não são atribuídas à proteína, tais como a altura do pico de densidade eletrônica e a posição da molécula de água em relação aos átomos da proteína, com os quais poderá

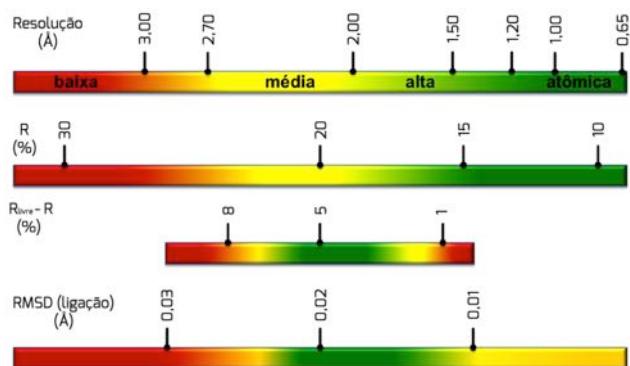



Figura 17-13: Critérios sugeridos para avaliação da qualidade de modelos de estruturas cristalográficas de macromoléculas, de adequado (verde) a inadequado (vermelho). Diferença entre o Rlivre e Rfator > 7% indica baixa correlação entre os dados experimentais e o modelo estrutural. Entretanto, se essa diferença for < 2% sugere-se que o conjunto de dados esteja demasiadamente "preso". Valores de RMSD (ver capítulo 8) indicam a presença de erros no modelo. Por outro lado, valores excessivamente baixos de RMSD (por exemplo, 0,004 Å) indicam excesso nas restrições estereoquímicas, com maior peso à otimização da geometria em detrimento dos dados de difração experimental durante os ciclos de refinamento.

formar ligações de hidrogênio.

Frequentemente, densidades eletrônicas próximas à cadeia polipeptídica são atribuídas a íons provenientes das soluções de cristalização, como sódio, cálcio e amônio. Em geral, essas densidades apresentam características específicas como formas, estado de coordenação ou propriedades eletrônicas que auxiliam a identificação correta do íon e o seu modo de ligação.

O número de moléculas de águas que podem ser identificadas e associadas a um determinado modelo estrutural irá depender da qualidade do modelo e dos dados cristalográficoa (ou seja, da sua resolução). Por exemplo, em estruturas de média resolução (2,5 a 3,0 Å) o número de moléculas de água esperado é baixo, pois apenas aquelas moléculas que estão fortemente associadas à proteína (usualmente localizadas no sítio ativo ou em outras regiões funcionais) podem ser corretamente posicionadas.

Já em estruturas de alta resolução (1,0–2,0 Å), pode-se identificar um número significativo de moléculas de água na superfície da proteína com boa precisão. Contudo, é importante mencionar que a utilização de moléculas de água em demasia em um modelo final pode mascarar regiões da densidade eletrônica e induzir a erros de interpretação, como a atribuição de águas a densidades que correspondem a cadeias laterais dos resíduos, outros tipos de solventes ou ligantes.

Como o Rfator pode ser interpretado como uma medida de quanto a densidade eletrônica é satisfeita, moléculas de água mal posicionadas podem diminuir o valor para o Rfator, porém, sem melhorar a acurácia do modelo. Nesses casos, a comparação entre os valores de Rfator e Rlivre é fundamental para avaliar a possibilidade de sobreajuste do modelo (diferença entre Rlivre e Rfator > 7%). A Tabela 4-13 apresenta valores representativos das estatísticas de refinamento para um bom modelo cristalográfico.

Uma estratégia frequentemente empregada para a identificação de erros de interpretação em modelos estruturais baseia-se nas características geométricas dos aminoácidos e das estruturas 2^{árias} (como distâncias, ângulos de ligação e diedros φ e ψ, ver capítulo 2).

As distâncias interatômicas e ângulos de ligação dos resíduos de aminoácidos são bem conhecidos e empregados como guia para avaliação de modelos estruturais. A medida é expressa pelo valor de RMSD para todas as distâncias e ângulos de ligação na proteína em estudo.

As relações entre os ângulos diedrais para os átomos da cadeia principal que contém estrutura 2^{ária} foram analisadas em termos de valores permitidos e proibidos em um gráfico conhecido como Gráfico de Ramachandran (Figura 18-13, ver capítulo 2).

Contudo, faz-se necessário salientar que alguns resíduos podem localizar-se fora das regiões permitidas por diferentes razões. Por exemplo, o resíduo de glicina, devido à ausência de uma cadeia lateral volumosa, pode ser encontrado fora das regiões permitidas. Por outro lado, o resíduo de prolina pode localizar-se em regiões proibidas em função de isomeria estrutural (isto é, isômeros *cis* e

Tabela 4-13: Exemplo de estatísticas de refinamento de uma estrutura de boa qualidade. Dados referentes aos estudos cristalográficos para a determinação da estrutura celobiohidrolase I de *Trichoderma harzianum* (PDB ID 2YOK).

Refinamento	
Resolução	45,3-1,67 (1,71-1,67)
Rfator/Rlivre (%)	14,6/17,3
Número de átomos	
Proteína	3193
N-acetil-D-GlcN	42
PEG	23
Água	562
Fator B (Å2)	
Proteína	10,3
N-acetil-D-GlcN	29,7
PEG	30,4
Água	24,2
RMSD	
Tamanho de ligação (Å)	0,011
Ângulo de ligação (º)	1,331

trans).

Ocasionalmente, se a resolução for alta o suficiente para permitir uma interpretação precisa, um resíduo pode aparecer fora dos limites aceitáveis (Figura 18-13). Exemplos como esse não são incomuns e, portanto, é fortemente recomendada a inspeção criteriosa de todos os resíduos de uma proteína, principalmente aqueles indicados em regiões não favoráveis no gráfico de Ramachandran.

Planejamento baseado na estrutura do receptor

Os avanços nas ciências biomédicas vem contribuindo significativamente para a identificação e validação de novos alvos moleculares de interesse terapêutico. Além disso, iniciativas como os programas genoma e proteoma de vários organismos têm fornecido dados importantes para o detalhamento das bases moleculares responsáveis pela estrutura e função de biomoléculas.

Simultaneamente, o aprimoramento das técnicas de determinação estrutural e análise de moléculas, como a cristalografia de raios-X, ressonância magnética nuclear (RMN) e a calorimetria, têm contribuído substancialmente para a melhor compreensão dos componentes energéticos e espaciais que compõem as interações entre fármacos e receptores.

Nas últimas décadas, os métodos cristalográficos ganharam enorme destaque como estratégia útil para o planejamento de fármacos. A sua aplicação vai desde os estudos em pesquisa básica, visando à elucidação das características estruturais e funcionais de alvos moleculares, até a pesquisa aplicada, caracterizada pela aplicação do conhecimento estrutural para a identificação de moléculas com atividade biológica e otimização de propriedades farmacodinâmicas e farmacocinéticas.

Atualmente, um dos maiores desafios na área de planejamento de novos fármacos é aumentar a taxa de sucesso na identificação de novas entidades químicas (NCEs, *new*

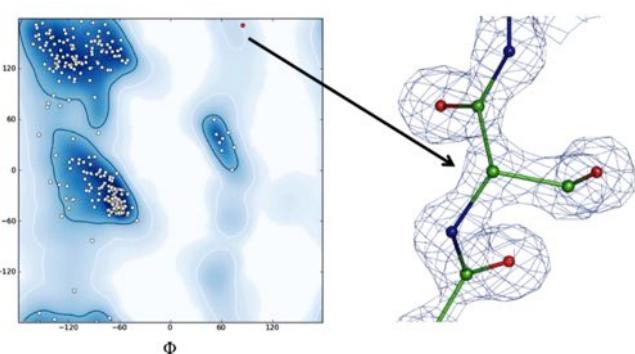



Figura 18-13: Gráfico de Ramachandran representativo para uma estrutura de boa qualidade. Destaque para o resíduo de serina que, apesar de localizado em um uma região proibida, é perfeitamente corroborado pelo mapa de densidade eletrônica.

chemical entities). Nesse contexto, destaca-se a estratégia de grande impacto denominada planejamento baseado na estrutura do receptor (SBDD, *Structure Based Drug Design*). Os métodos de SBDD se baseiam no conhecimento da informação 3D da macromolécula alvo, que geralmente é obtida de estruturas determinadas por cristalografia de raios-X, por RMN ou através de modelagem por homologia.

As estratégias de SBDD têm como princípio o entendimento do mecanismo que leva ao aparecimento de doenças, aliado à identificação de alvos moleculares que forneçam novas oportunidades para o desenvolvimento de NCEs. O planejamento de fármacos utilizando estruturas 3D de biomoléculas proporcionou o desenvolvimento de uma importante variedade de inovações terapêuticas, trazendo benefícios notáveis à saúde humana das mais diversas populações mundiais.

A informação sobre o modo de ligação de substâncias bioativas, levando em conta a complementaridade de interações entre ligante e receptor, é de grande utilidade no planejamento de candidatos a novos fármacos. A partir da obtenção e avaliação farmacológica de séries de compostos sintéticos, pode-se estudar a relação entre as suas diferenças estruturais e as atividades medidas (relação estrutura atividade), estabelecendo pressupostos úteis na elaboração de estratégias de modificação molecular.

Devido à complexidade e à quantidade de informação gerada, métodos de modelagem molecular (como ancoramento, modelagem comparativa e dinâmica molecular, vistos em capítulos anteriores) são constantemente empregados para caracterizar as interações predominantes entre ligantes e receptores biológicos. Os compostos bioativos mais promissores nas diversas etapas de investigação podem ser então submetidos a ensaios cristalográficos, visando tanto validar os resultados computacionais quanto refinar e ampliar o nível de informação molecular. Um dos principais exemplos de doenças que se beneficiaram destas técnicas envolve o tratamento da AIDS, causada pelo vírus da imunodeficiência humana (HIV).

Devido à função central exercida no desenvolvimento do vírus, a protease do HIV tornou-se um alvo prioritário de muitas indústrias farmacêuticas. As primeiras investigações para a identificação de inibidores da protease de HIV se basearam em dados estruturais de um modelo teórico construído com o auxílio de métodos de modelagem comparativa. A primeira estrutura cristalográfica da protease de HIV foi resolvida em sua forma nativa no final da década de 1980. Subsequentemente, mais de 250 complexos entre inibidores e esta protease foram obtidos, fornecendo bases estruturais sólidas para o desenvolvimento de uma série de fármacos, ainda em uso terapêutico.

O planejamento de inibidores da protease de HIV é um dos exemplos de maior sucesso na aplicação dos métodos experimentais e computacionais ao desenvolvimento de novos fármacos. O desenvolvimento do peptideomimético saquinavir (Invirase®, Roche), primeiro inibidor da protease de HIV aprovado pelo FDA (*Food and Drug Administration*) nos Estados Unidos para o tratamento da AIDS, em 1995, teve sua origem em dados cristalográficos obtidos com os inibidores peptídeos desta protease (Figura 19-13).

Os modelos de interação, obtidos por cristalografia, indicavam que a substituição isostérica da ligação amídica central por um grupo hidroxietilamina estaria relacionada com o aumento de potência e seletividade. Isto motivou a síntese e avaliação bioquímica de uma série de análogos, que confirmaram esta hipótese.

A etapa seguinte dos estudos consistiu na avaliação do tamanho da sequência peptídica para uma ótima inibição. Estudos de modelagem molecular foram empregados para priorizar a síntese de derivados com tamanhos distintos de cadeia. Aliados a testes biológicos, estes experimentos mostraram que o tamanho mínimo da cadeia peptídica deveria ser de 5 resíduos de aminoácidos.

Em seguida, foi investigada a influência da variação das cadeias laterais nas unidades peptídicas. Vários análogos foram obtidos, embora nenhum tenha apresentado melhora considerável da potência inibitória. Por outro lado, a substituição do resíduo de prolina na

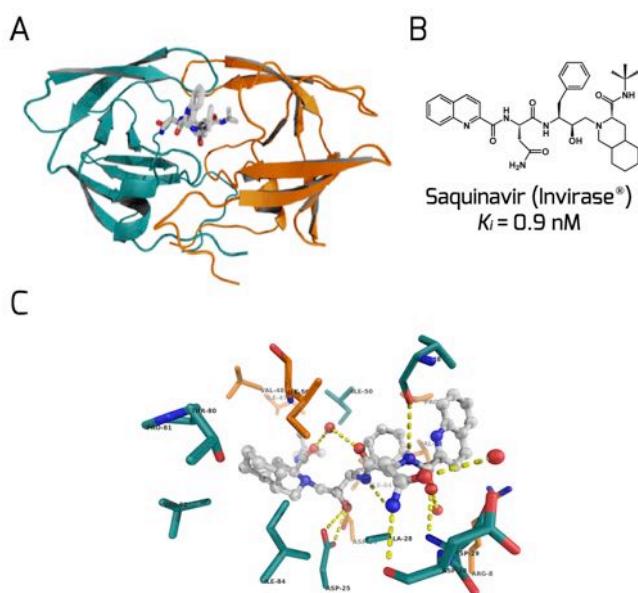



Figura 19-13: (A) Homodímero da protease de HIV-1 em complexo com inibidor saquinavir (PDB ID 1FB7). (B) Estrutura química do saquinavir. (C) Detalhes do modo de ligação do inibidor saquinavir no sítio ativo da enzima.

molécula do inibidor por grupos piperidina ou 3-carbonil-decahidro-isoquinolina (DIQ) acarretou em uma melhora significativa da potência inibitória.

Os modelos de interação sugeriram que a maior potência do derivado DIQ (saquinavir, Figura 19B-13) estaria relacionada a um menor grau de liberdade conformacional conferido por este substituinte, indicando um favorecimento entrópico para a energia livre de ligação. Posteriormente, a análise do complexo cristalográfico saquinavir-protease revelou que a porção DIQ do inibidor adotava uma conformação de energia mínima, característica de grupos cíclicos saturados, confirmando o modo de ligação predito (Figura 19C-13).

As informações obtidas no desenvolvimento do saquinavir serviram de base para o planejamento de novos inibidores da protease de HIV, tais como ritonavir (Norvir®, Abbott), indinavir (Crixivan®, Merck Sharp & Dohme) e nelfinavir (Viracept®, Agouron Phamaceuticals).

Genoma estrutural

Os sucessos conquistados pelos projetos genômicos deram um importante suporte à abordagem do tipo "larga escala" na atividade científica. No campo da cristalografia, as ideias genômicas foram extrapoladas procurando retornar à sociedade um conjunto de informações representativas da biodiversidade do universo proteico, gerando estruturas tridimensionais em nível atômico para a maior parte das proteínas facilmente obtidas à partir do conhecimento de suas sequências de DNA (www.nigms.nih.gov / Initiatives / PSI.htm).

A escala dessa abordagem é estabelecida, inicialmente, na definição e seleção de sequências de aminoácidos mais susceptíveis à determinação estrutural, procurando-se evitar proteínas mais "problemáticas".

Contudo, o esforço empregado na determinação do genoma estrutural é significativamente maior do que no sequenciamento. Isto se deve à grande diferença de complexidade dos métodos envolvidos e à variabilidade no comportamento dos alvos proteicos em diferentes estágios do processo de determinação estrutural em larga escala.

Uma vez que a estrutura tridimensional de uma proteína é muito mais conservada que sua sequência de aminoácidos, o conhecimento de seu enovelamento torna-se uma ferramenta muito valiosa para se estudar e descobrir relações evolucionárias imperceptíveis em nível de sequência. Essas similaridades estruturais podem, por exemplo, sugerir propriedades funcionais às proteínas de funções ainda desconhecidas.

A contribuição mais prontamente visível da genômica estrutural é a rápida expansão do número de estruturas de proteínas disponíveis no PDB e, geralmente, a um custo reduzido devido à eficiência e otimização das técnicas desenvolvidas em centros especializados.

Uma seleção adequada de alvos é fundamental para assegurar que as estruturas resolvidas por esses centros sejam realmente valiosas para toda a comunidade científica e industrial, seja devido ao interesse intrínseco das proteínas estudadas, ou visando uma melhoria do mapeamento do universo proteico, fornecendo modelos para novos estudos de modelagem comparativa (Figura 20-13).

Nesse contexto, uma segunda contri-

buição importante dos projetos de genômica estrutural para a comunidade científica é o desenvolvimento de métodos e tecnologias para a produção eficiente de proteínas e determinação estrutural, que possam ser adotados em laboratórios de pesquisa menores contribuindo, assim, com o avanço da área ao retor do mundo.

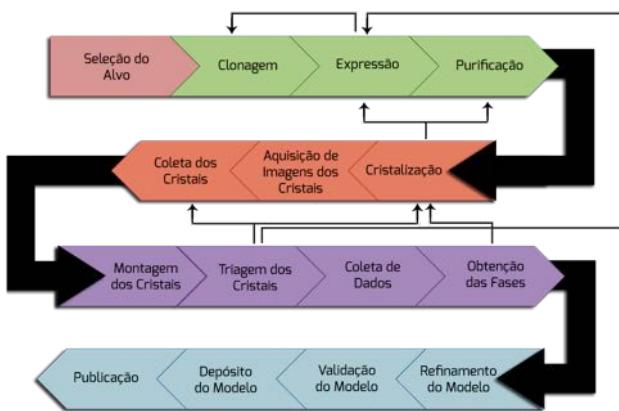



Figura 20-13: Fluxograma representativo de um projeto de genoma estrutural.

13.7. Conceitos-chave

Cristal: sólido no qual os átomos constituintes estão organizados num padrão tridimensional bem definido, que se repete no espaço, formando uma estrutura com uma geometria específica.

Cristalização: processo de separação sólido-líquido no qual há transferência de massa de um soluto a partir de uma solução líquida supersaturada para uma fase sólida cristalina pura.

Cromatografia: método de separação e identificação dos componentes em uma mistura. Ampalmente empregado para a purificação de proteínas.

Difração: fenômeno de interação entre a radiação eletromagnética com a matéria com consequente dispersão dessa radiação.

Expressão em sistema heterólogo: expressão de um gene (ou parte dele) em um organismo hospedeiro, o qual naturalmente não possui este gene (ou fragmento de gene).

Luz síncrotron: acelerador de partículas poligonal que produz luz usando eletroímãs poderosos e ondas de radiofrequência para acelerar elétrons a uma velocidade próxima à da luz em um anel de armazenamento.

Mapa de densidade eletrônica: Região de maior probabilidade de se encontrar os elétrons. O mapa de densidade eletrônica é o resultado final de um experimento de difração de raios-X. A análise detalhada do mapa orienta a construção do modelo estrutural da proteína.

Padrão de difração: padrão produzido a partir de uma estrutura tridimensional periódica, como átomos de um cristal, que contém informação sobre a separação dos planos cristalográficos. A análise do padrão de difração permite que se possa deduzir a estrutura do cristal.

PDB: banco de dados de proteínas de acesso livre em http://www.rcsb.org.

Raios-X: radiação eletromagnética com comprimento de onda entre 0,01-10 nm (0,1-100 Å).

Refinamento: processo supervisionado de construção e ajuste do modelo estrutural aos dados de difração de raios-X.

Sistema de clonagem LIC: estratégia em biologia molecular para a clonagem independente de ligação capaz de aumentar a taxa de sucesso na obtenção de proteína expressa na forma solúvel, com alta pureza e em grande quantidade.

Solução de cristalização: solução que favorece a cristalização de proteínas constituída de componentes como agentes tamponantes, aditivos que facilitam o processo de cristalização e agentes precipitantes.

13.8. Leitura recomendada

BERGFORS, T. ***Protein Crystallization***. 2nd.ed. San Diego: International University Line, 2009.

BLUNDELL, T. L.; JOHNSON, L. N. ***Protein Crystallography***, 1st.ed. Academic Press, 1976.

JANSON, J.-C. ***Protein Purification: Principles, High Resolution Methods, and Applications***. 3rd.ed. New Jersey: Wiley, 2011.

MCPHERSON, A. ***Introduction to Macromolecular Crystallography***. Hoboken: John Wiley & Sons, 2009.

RUPP, B. ***Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology***. New York: Garland Science, 2010.

STOUT, G. H.; JENSEN, L. H. ***X-ray Structure Determination: A Practical Guide***. John Wiley & Sons, 1989.

WLODAWER, A.; et al. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. ***FEBS j.*** 275, 1–21, 2008.